

A GENERAL FRAMEWORK FOR LOCATING HYPERPLANES TO FITTING SET OF POINTS

VÍCTOR BLANCO, JUSTO PUERTO, AND ROMÁN SALMERÓN

ABSTRACT. This paper presents a family of new methods for locating/fitting hyperplanes with respect to a given set of points. We introduce a general framework for a family of aggregation criteria of different distance-based errors. The most popular methods found in the specialized literature can be cast within this family as particular choices of the errors and the aggregation criteria. Mathematical programming formulations for these methods are stated and some interesting cases are analyzed. It is also proposed a new goodness of fitting index which extends the classical coefficient of determination. A series of illustrative examples and extensive computational experiments implemented in R are provided to show the performances of some of the proposed methods.

1. INTRODUCTION

The problem of locating hyperplanes with respect to a given set of point is well-known in Location Analysis [41]. This problem is closely related to another common question in Data Analysis: to study the behavior of a given set of data with respect to a fitting body expressed with an equation of the form $f(X_1, \dots, X_d) = 0$. This last problem reduces to the estimation of the ‘best’ function f that expresses the relationship between the provided data or in other words to the location of the surface $f(X) = 0$ that minimizes some aggregation function of the distances of the points (data set) to the dimensional facility f (see [1, 14, 15]). In many cases and for the sake of simplicity, the family of functions where f belongs to is usually fixed and then, real parameters of such a function must be determined. The most widely used family of functions considered in this framework, probably because of its simplicity, is the family of linear functions, namely the above equation is of the form $f(X_1, \dots, X_d) = \beta_0 + \sum_{k=1}^d \beta_k X_k = 0$ for $\beta_0, \beta_1, \dots, \beta_d \in \mathbb{R}$.

To perform such a fitting, we are given a set of points $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$, and one tries to find the values $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_d)$ that minimize some measure of the deviation of the data with respect to the hyperplane $\mathcal{H}(\hat{\beta}) = \{z \in \mathbb{R}^d : \hat{\beta}_0 + \sum_{k=1}^d \hat{\beta}_k z_k = 0\}$. For a certain observation $x \in \mathbb{R}^d$ in the data set, such a deviation is usually known as the residual (terminology borrowed from the Statistical Regression literature). In a general framework, for a given point $x \in \mathbb{R}^d$, we define the *residual* of a model as a mapping $\varepsilon_x : \mathbb{R}^{d+1} \rightarrow \mathbb{R}_+$, that maps any set of coefficients $\beta = (\beta_0, \dots, \beta_d) \in \mathbb{R}^{d+1}$, into a measure $\varepsilon_x(\beta)$ that represents the deviation of the given point x from the hyperplane with those parameters. The larger this measure, the worse the fitting for such a point x . The final goal of fitting an hyperplane for a given set of points $\{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ is to find the coefficients minimizing a globalizing function, $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$, of the residuals of all the points. Equivalently, the fitting problem consists in locating a hyperplane minimizing the globalizing function Φ of the distances from the demand points to the hyperplane. Different choices for the residuals and the globalizing criteria will give, in general, different optimal values for the parameters and thus different properties for the resulting hyperplanes. This problem is not new and some of these fitting criteria, as the minisum, minimax and some other robust versions, have been analyzed from a Locational analysis perspective (see [12, 26, 38, 39, 40, 41], among other).

The most natural approach to locate a hyperplane is to consider that residuals, with respect to given points, are individual measures of error and thus, each residual should be minimized independently of the remaining. Obviously, this approach gives rise to a multicriteria problem [11, 30]. It is clear that this simultaneous minimization will not be possible in most of the cases and then several strategies can be followed: one can try to find the set of Pareto fitting curves [11] or alternatively, to apply an aggregation function that incorporates the holistic preference of the Decision-Maker on the different residuals. This last choice is very difficult and the usual approach is to apply the principle of complete uncertainty leading to additive aggregations.

The most popular methods to compute the coefficients of an optimal hyperplane consider that the residuals are the differences from one of the coordinates of the space (which are usually known as vertical/horizontal distances). In this paper we present a new framework for optimally locating/fitting hyperplanes to a set of points that allows the decision-maker to decide within a wide family of residuals and criteria which is the “best” for a given sample of data. One of the main contributions of our proposal is the use of modern mathematical programming tools to solve the problems which are involved in the computation of the parameters of the fitting models. The optimization models for those problems range from continuous convex programming (CP) to mixed

2010 *Mathematics Subject Classification.* 90B85 and 90C26 and 52C35 and 65D10..

Key words and phrases. Fitting Hyperplanes and Mathematical Programming and Location of Structures and Robust Fitting.

integer nonlinear programming (MINLP) through linear programming (LP). Many of the formulations described in this paper have been implemented in R in order to be available for data analysts.

The framework in this paper introduces a family of combinations residuals-criteria that allows a great flexibility to accommodate hyperplanes to set of points [32, 25]. This new framework can be easily combined with some of the mathematical programming techniques for feature selection, to “choose” a fixed number of coordinates to explain the dependence between the different dimensions [7], with classification schemes [6], or when the coefficients of the linear manifold are required to fulfill a set of linear equations/inequalities. This framework can also accommodate general forms of regularization, as upper bound on the ℓ_2 -norm of the coefficients [21], since it would only mean to add additional constraints to the mathematical programming formulations proposed in the paper. The complexity of solving the resulting model depending on the difficulty of the considered regularization constraints.

In order to compare the *goodness of the fitting* for the different models we have developed a new generalized measure of fit. This task becomes difficult when one tries to compare fitting hyperplanes which are built based on different paradigms and purposes. The new measure is provided in order to make meaningful comparisons. This proposal is based on a generalization of the classical coefficient of determination, that will allow to measure how good is an optimal hyperplane with respect to the best constant model, $X_d = \beta_0$. This measure will extend the standard coefficient of determination for least squares fitting. We also perform an extensive series of experiments to validate the application of our results applied with different objectives to several set of data.

In our framework, errors are measured as shortest distances, based on a norm, between the given points and the fitting surface. This makes the location problem geometrically invariant which is an interesting advance with respect to vertical/horizontal residuals. Through the paper we observe that this framework also subsumes as particular cases the standard location methods that consider residuals based on vertical distances (commonly used in Statistics); as well as most of the particular cases of fitting linear bodies using vertical distances but different aggregation criteria described in the literature, as ℓ_p fitting (ℓ_p -norm criterion), least quantile of squares [36, 7], least trimmed sum of squares [35, 3], etc. As previously mentioned, the problem of optimally locating an hyperplane with respect to a set of demand points is closely related to the estimation phase in multivariate linear regression, where several methods have already been proposed. However, the use of nonstandard residuals is not usual in the literature of regression analysis although orthogonal (ℓ_2) residuals have been already used, see e.g. Euclidean Fitting [5, 13, 34] or Total Least Squares [45], mainly applied to bidimensional data. Quoting the reasons for that fact given by Giloni and Padberg in [19]: “we have left out a summary of linear regression models using the more general ℓ_τ -norms with $\tau \notin \{1, 2, \infty\}$ for which the computational requirements are considerably more burdensome than in the linear programming case (as they generally require methods from convex programming where machine computations are far more limited today).”

The paper is organized as follows. In Section 2 we introduce the new framework for fitting hyperplanes as well as some results that allows to interpret the results for practical purposes. Next, a residual-aggregation dependent goodness of fitting index is defined and it is presented an efficient approach for its computation. Section 3 is devoted to the analysis of the classical location methods under the new framework, more precisely, mathematical programming models for adequate aggregation criteria and residuals are provided for: 1) least sum of squares; 2) least absolute deviation; 3) least quantile of squares and 4) least trimmed of squares fitting. In Sections 4 and 5 we present new methods for the location of hyperplanes assuming that the residuals are measured as the smallest norm-based distance between the given points (data set) and the linear fitting body using polyhedral norms (Section 4) and ℓ_τ norms (Section 5), respectively. We also present, in Section 5, outer an inner approximations for solving the resulting MINLP problems for ℓ_p -norms residuals. Finally, Section 6 is devoted to the computational experiments. We report results for synthetic data and for the classical data set given in [16].

2. A FLEXIBLE METHODOLOGY FOR THE LOCATION OF HYPERPLANES

Given is a set of n observations or demand points (depending that we use the *jargon* of data analysis or location analysis, respectively) in a $(d + 1)$ -dimensional space, $\{x_1, \dots, x_n\} \subset \mathbb{R}^{d+1}$ (we will assume, for a clearer description of the models, that the first, the 0 – th, component of x_i is the one that account for the intercept in the model, being $x_{10} = \dots = x_{n0} = 1$). Next, we analyze ways of fitting these observations to a linear form (hyperplane). For any $y \in \mathbb{R}^{d+1}$, we shall denote $y_{-0} = (y_1, \dots, y_d)$, i.e. the vector with the last d coordinates of y excluding the first one. We consider here a flexible framework for the problem of locating/fitting hyperplanes that includes as special cases the classical and most modern models found in the specialized literature. First, we assume that the point-to-hyperplane deviation is modelled by a residual mapping $\varepsilon_x : \mathbb{R}^{d+1} \rightarrow \mathbb{R}_+$, $\varepsilon_x(\beta) = D(x_{-0}, \mathcal{H}(\beta))$, being D a distance measure in \mathbb{R}^d . This residual represents how “far” is the point (observation) $x \in \mathbb{R}^{d+1}$ with respect to the hyperplane $\mathcal{H}(\beta) = \{y \in \mathbb{R}^d : (1, y^t)\beta = 0\}$ (Some times we will write the hyperplane as $\beta^t X = 0$, with $\beta = (\beta_0, \beta_1, \dots, \beta_d)^t \in \mathbb{R}^{d+1}$.)

Furthermore, the residuals for each demand point are aggregated using a globalizing function $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$, which for a set of residuals $\varepsilon_1, \dots, \varepsilon_n$ gives an overall measure of the deviations of the whole data set with

respect to the hyperplane. With this setting, ones tries to minimize such a globalizing measure of the residuals with respect to all the given demand points.

With this notation, the *Fitting Hyperplane Problem* (FHP) consists in finding $\hat{\beta} \in \mathbb{R}^{d+1}$ such that:

$$(FHP(\Phi, \varepsilon)) \quad \hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^{d+1}} \Phi(\varepsilon_x(\beta)),$$

where $\varepsilon_x(\beta) = (\varepsilon_{x_1}(\beta), \dots, \varepsilon_{x_n}(\beta))^t$ is the vector of residuals.

Note that the difficulty of solving FHP(Φ, ε) depends of the expressions for the residuals and the aggregation criterion Φ . If Φ and ε_x are linear, the above problem becomes a linear programming problem. In this paper, we consider a general family of aggregation criteria that includes as particular cases most of the classical ones used in the literature. Some of those criteria have been already considered for the sake of outlier detection [37, 48] or as robust alternatives to the standard linear regression approach [7, 19].

Let $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ and let $\varepsilon \in \mathbb{R}^n$ be the vector of residuals of all of the demand points in the given data set. We consider aggregation criteria $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}_+$ defined as:

$$(1) \quad \Phi(\varepsilon) = \sum_{i=1}^n \lambda_i \varepsilon_{(i)}^p$$

where $\varepsilon_{(i)} \in \{\varepsilon_1, \dots, \varepsilon_n\}$ is such that $\varepsilon_{(1)} \leq \dots \leq \varepsilon_{(n)}$. Observe that this operator defines a multiparametric family (called *ordered median functions* [32]) that depending on the choice of the λ -weights captures many of the models proposed in the literature.

Note that the above shape of Φ is symmetric and, for non negative lambda weights, a monotone function that ensures that the ordering of the individual residuals do not affect the overall goodness of the fitting. Moreover, it also implies that a componentwise smaller vector of residuals gives rise to a more accurate fitting.

The natural implication of the assumption made about the definition of residuals is that, as expected, the response (projection) of a demand points on a given hyperplane differs from the classical evaluation and it must be the closest point, with respect to the distance D, in the located hyperplane $\mathcal{H}(\beta)$.

Lemma 1. *For a given point $z^t = (1, z_1, \dots, z_d)$ and the hyperplane $\mathcal{H}(\beta)$ the response \hat{z} consistent with the residual $\varepsilon_z = \min_{y \in \mathcal{H}(\beta)} \|z_{-0} - y\|$ is given by*

$$\hat{z} = z_{-0} - \frac{\beta^t z}{\|\beta_{-0}\|^*} k(\beta),$$

where $\|\cdot\|^*$ is the dual norm to $\|\cdot\|$ and $k(\beta) = \arg \max_{\|x\|=1} \beta_{-0}^t x$. Moreover,

$$(2) \quad \varepsilon_z = \frac{|\beta^t z|}{\|\beta_{-0}\|^*}.$$

Proof. The proof follows applying [24, Theorem 2.1] to the definition of residual $\varepsilon_z = \min_{y \in \mathcal{H}(\beta)} \|z_{-0} - y\|$. \square

From the above result, the response for a point with a unknown coordinate (w.l.o.g, the last component, d), namely $z = (1, z_1, \dots, z_{d-1}, 0)^t$, will be given by:

$$\hat{z}_d = -\frac{\beta^t z}{\|\beta_{-0}\|^*} k(\beta)_d.$$

Hence, differentiating \hat{z} with respect to each z_j , $j = 1, \dots, d-1$, we get

$$\frac{\partial \hat{z}_d}{\partial z_j} = -\frac{\beta_j}{\|\beta_{-0}\|^*} k(\beta)_d,$$

which may be interpreted as the marginal variation of the d -th coordinate with respect to j -th coordinate whenever the other dimensions remain constant.

Explicit expressions for such projections, namely, ℓ_1 , ℓ_∞ and ℓ_τ -norms, for $\tau > 1$ are described in the following lemma.

Lemma 2. *Let $z = (1, z_1, \dots, z_d)^t$, then*

(1) *If D is the ℓ_1 - distance,*

$$\hat{z}_k = \begin{cases} z_k & \text{if } |\beta_k| \neq \max\{|\beta_j| : j = 1, \dots, d\}, \\ z_k - \frac{\beta^t z}{\|\beta_{-0}\|_\infty} v_k, & \text{if } \beta_k = \max\{|\beta_j| : j = 1, \dots, d\}, \\ z_k + \frac{\beta^t z}{\|\beta_{-0}\|_\infty} v_k, & \text{if } \beta_k = -\max\{|\beta_j| : j = 1, \dots, d\}, \end{cases}$$

for $k = 1, \dots, d$, and for some $v_1, \dots, v_d \geq 0$ such that $\sum_j v_j = 1$.

(2) If D is the ℓ_∞ - distance,

$$\hat{z}_k = \begin{cases} z_k - \frac{\beta^t z}{\|\beta_{-0}\|_1}, & \text{if } \beta_k > 0, \\ z_k + \frac{\beta^t z}{\|\beta_{-0}\|_1}, & \text{if } \beta_k < 0, \end{cases} \quad k = 1, \dots, d.$$

(3) If D is the ℓ_τ - distance with $1 < \tau < +\infty$ then

$$\hat{z}_k = z_k - \frac{\beta^t z}{\|\beta_{-0}\|_\nu} k_\tau(\beta)_k, \quad k = 1, \dots, d$$

and

$$k_\tau(\beta)_k = \begin{cases} \frac{\text{sign}(\beta_k) |\beta_k|^{\nu/\tau}}{(\sum_{j=1}^d |\beta_j|^\nu)^{1/\tau}} & \text{if } \beta_k \neq 0 \\ 0 & \text{if } \beta_k = 0, \end{cases} \quad k = 1, \dots, d,$$

being ν such that $\frac{1}{\tau} + \frac{1}{\nu} = 1$.

Proof. The proof of items 1. and 2. can be found in [24]. The proof of item 3. follows from the Lagrangian optimality condition applied to $\max_{\|z\|_\tau=1} \beta_{-0} z$. First, we observe that a Lagrange multiplier exists since the problem is regular at any point of the ℓ_τ unit ball. Next, the Lagrangian function is $L(z, \lambda) = \beta_{-0} z - \lambda \sum_{k=1}^d |z_k|^\tau$. Therefore, its partial derivatives are: $\frac{\partial L}{\partial z_k} = \beta_k - \lambda \tau |z_k|^{\tau-1} \text{sign}(z_k)$, for all $k = 1, \dots, d$. Hence, equating to zero the partial derivative, it follows that for any index k such that $z_k^* \neq 0$

$$(3) \quad \lambda^* = \frac{\beta_k}{\tau |z_k^*|^{\tau-1}} \text{sign}(z_k^*).$$

Let us define the sets $I = \{k : \beta_k > 0\}$, $J = \{k : \beta_k < 0\}$, $K = \{k : \beta_k = 0\}$. Now from equation (3), and taking into account that $\|z\|_\tau = 1$, we obtain:

$$|z_k^*|^\tau = \begin{cases} \frac{(\text{sign}(z_k^*) \beta_k)^\nu}{(\sum_{j=1}^d \text{sign}(z_j^*) \beta_j)^\nu} & \text{if } k \in I \cup J \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, the hessian of L is diagonal and all its entries are negative, namely $\frac{\partial^2 L}{\partial z_k^2} = -\lambda \tau(\tau-1) |z_k^*|^{\tau-2}$. This implies that z^* and λ^* are local maxima.

In the particular case of $\tau = 2$ then one can check that $k_2(\beta)_k = \beta_k$ which simplifies the above expression. \square

We note in passing that $\varepsilon_x = D_{\|\cdot\|}(x_{-0}, \mathcal{H}(\beta))$ and thus, according to the Lemma 1

$$(4) \quad D_{\|\cdot\|}(x_{-0}, \mathcal{H}) = \frac{|\beta^t x|}{\|\beta_{-0}\|_*}.$$

Observe also that when the demand points in the data set lie exactly on the hyperplane \mathcal{H} all the proposed methods FHP(Φ, ε) determine the same hyperplane \mathcal{H} as an optimal fitting, for any norm-based residuals while using vertical distance residuals will never produce hyperplanes in the form $\mathcal{H} = \{z \in \mathbb{R}^d : \beta_0 + \beta_1 z_1 + \dots + \beta_{d-1} z_{d-1} = 0\}$ since the “traditional” methods do not allow zero coefficients for the *dependent* coordinate. Note also that the vertical distance based methods assume that errors are present only in one of the components (the so-called dependent), so the rest of the variables should be error-free. In the proposed general framework, this is no longer assumed since there is no distinction between dependent and independent variables for the location/fitting procedure, so errors may be considered in all the components of the points in the given data set.

Remark that the standard residual (vertical distance) is a distance measure that is not induced by a norm, but its expression can be written in an analogous form and so it fits to the shape of the distances that are considered in this paper. In particular, the vertical distance (with respect to the last coordinate) may be defined as:

$$(5) \quad D_V(x, H) = \frac{\left| \beta_d x_d - \sum_{i=1}^{d-1} \beta_i x_i - \beta_0 \right|}{|\beta_d|}.$$

The above aggregation criteria (1) and residual functions (2) are rather general and exhibit good structural properties. On the one hand, they accommodate most of the already considered fitting methods in the literature. On the other hand, one can always exploit its properties and different representations in order to solve the optimization problem FHP(Φ, ε). In the following we prove some structural properties that imply some sources of solvability of the problem on hands.

For the sake of completeness, we recall the concept of *difference of convex* (D.C.) function. A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be a D.C. if there exist $g, h : \mathbb{R}^d \rightarrow \mathbb{R}$ convex functions such that f can be decomposed as

the difference between g and h . Optimization problems where the objective function and/or the constraints are defined by D.C. functions are called D.C. programming problems and they play an important role in nonconvex optimization because of its theoretical aspects as well as its wide range of applications (see [44]).

Lemma 3. *The globalizing function $\Phi(\varepsilon_x(\beta))$ is a D.C. function.*

Proof. In order to prove that the function Φ is D.C. we will find a convenient representation where we can apply properties of the algebra of D.C. functions. To this for, we introduce the functions:

$$\varphi_r(\beta) := \min \left\{ \max \{ \varepsilon_{x_{i_1}}(\beta)^p, \dots, \varepsilon_{x_{i_r}}(\beta)^p : i_1 < i_2 < \dots < i_r, \forall i_1, i_2, \dots, i_r \} \right\},$$

for $r = 1, \dots, n$, where $\varepsilon_x(\beta) = D_{\|\cdot\|}(x_{-0}, \mathcal{H})$.

It is a simple observation that $\varphi_r(\beta)$ coincides with the p -power r -th residual sorted in non-decreasing sequence, namely $\varphi_r(\beta) = \varepsilon_{x_{(r)}}(\beta)^p$ for all $r = 1, \dots, n$. Hence, we get that $\Phi(\varepsilon_x(\beta)) = \sum_{i=1}^n \lambda_i \varphi_r(\beta)$.

To finish the proof it suffices to prove that each function φ_r is D.C. since linear combinations of D.C. functions are D.C.. Next, we start analyzing the residual function $\varepsilon_x(\beta) = d(x, \mathcal{H}(\beta))$. Assuming that d is a norm based distance given in the form of (4) or (5), one can use those expressions to conclude that for each observation x , $\varepsilon_x(\beta)$ is D.C. function of β . Raising to the power p with $p \geq 1$ is also D.C., since it is the result of composing with a convex function (observe that residuals are non-negative). Finally, the operations of taking maxima and minima of D.C. functions are closed within this family [44]. This proves that φ_r is D.C. for all r and this concludes the proof. \square

We note in passing that the D.C. character of our globalizing criterion allows the application of all the available results on the optimization of this class of functions (see e.g. [44]). In spite of that, we can give more efficient representations that may help latter in the resolution of particular hyperplanes. These representations are based on simpler functions which replace φ by more friendly classes of functions (with regards to the optimization phase).

Proposition 4. *The globalizing function $\Phi(\varepsilon_x(\beta)) := \sum_{i=1}^n \varepsilon_{x_i}(\beta)^p + \sum_{r=2}^n (\lambda_r - \lambda_{r-1}) \theta_r(\beta)$, where $\theta_r(\beta) = \max \left\{ \varepsilon_{x_{i_1}}(\beta)^p + \dots + \varepsilon_{x_{i_r}}(\beta)^p : \begin{matrix} \{i_1, \dots, i_r\} \subset \{1, \dots, n\} \\ i_1 < i_2 < \dots < i_r \end{matrix} \right\}$, $r = 2, \dots, n$. (The reader may observe that the functions θ_r are usually called r -centrum in the specialized literature of optimization ([32]).)*

Proof. This representation follows from the combination of the result in Lemma 3 and [20, Theorem 3.6]. \square

The following result states a mathematical programming formulation for the generalized fitting hyperplane problem, for any choice of Φ and ε_x .

Theorem 5. *Let $\{x_1, \dots, x_n\} \subseteq \mathbb{R}^{d+1}$ be a set of demand points, $\lambda \in \mathbb{R}_+^n$, $p = \frac{r}{s} \in \mathbb{Q}$ and $\|\cdot\|$ a norm in \mathbb{R}^d . The Problem FHP(Φ, ε) is equivalent to the following mathematical programming problem:*

$$\begin{aligned}
 (\text{LR}_{\Phi, \|\cdot\|}) \quad & \min \sum_{j=1}^n \lambda_j \theta_j \\
 (6) \quad & \text{s.t. } \varepsilon_i \geq \frac{|\beta^t x_i|}{\|\beta_{-0}\|^*}, & \forall i = 1, \dots, n, \\
 (7) \quad & z_i \leq \theta_j + M(1 - w_{ij}), & \forall i, j = 1, \dots, n, \\
 (8) \quad & z_i^s \geq \varepsilon_i^r, & \forall i = 1, \dots, n, \\
 (9) \quad & \sum_{i=1}^n w_{ij} = 1, & \forall j = 1, \dots, n, \\
 (10) \quad & \sum_{j=1}^n w_{ij} = 1, & \forall i = 1, \dots, n, \\
 (11) \quad & \theta_j \geq \theta_{j-1}, & \forall j = 2, \dots, n, \\
 & w_{ij} \in \{0, 1\}, & \forall i = 1, \dots, n, \\
 & z, \theta \in \mathbb{R}_+^n, \beta \in \mathbb{R}^{d+1}.
 \end{aligned}$$

Note that the above problem is a mixed integer non linear programming problem, whose continuous relaxation is in general non convex due to the constraints 6. Apart from the mathematical programming formulation above, one may use alternative (in some cases better) formulations for the ordering problems as those provided in [17]. In particular, some important special ordered median aggregation criteria allow to have a simpler formulation that avoids the use of binary variables. The following result shows a better formulation for the fitting problem under the assumption that $0 \leq \lambda_1 \leq \dots \leq \lambda_n$. We call this setting for lambda the *monotone case*.

Theorem 6. Let $\{x_1, \dots, x_n\} \subset \mathbb{R}^{d+1}$ be a set of demand points, $\lambda \in \mathbb{R}^n$, such that $0 \leq \lambda_1 \leq \dots \leq \lambda_n$, $p = \frac{r}{s} \in \mathbb{Q}$ with $r > s \in \mathbb{N}$, $\gcd(r, s) = 1$ and $\|\cdot\|$ a norm in \mathbb{R}^d . Then, $\text{FHP}(\Phi, \varepsilon)$ is equivalent to the following mathematical programming problem:

$$\begin{aligned} \min \quad & \sum_{j=1}^n v_j + \sum_{i=1}^n w_i \\ \text{s.t.} \quad & (6), (8), \\ & v_j + w_i \geq \lambda_i z_j, \forall i, j = 1, \dots, n, \\ & z_i, \theta_i \geq 0, v, w \in \mathbb{R}^n, \beta \in \mathbb{R}^{d+1}. \end{aligned}$$

Proof. The proof follows by the representation of the ordering between the residuals by permutation variables, which for $0 \leq \lambda_1 \leq \dots \leq \lambda_n$, allows to write the objective function in $\text{FHP}(\Phi, \varepsilon)$ as an assignment problem which is totally unimodular, so it can be equivalently rewritten using its dual problem. The interested reader is referred to [9] for further details on this transformation. \square

The reader may observe that, based on an alternative representation, the nonlinear constraints $z_i^s \geq \varepsilon_i^r$ for all $i = 1, \dots, n$ can be transformed into a set of second order cone constraints using the following result which is a simplified version of Lemma 1 in [9]. This implies that those constraints can be efficiently handled by nowadays nonlinear solvers since they are convex and friendly for the optimization.

Lemma 7. Let $r, s \in \mathbb{N} \setminus \{0\}$ with $\gcd(r, s) = 1$, and $k = \lfloor \log_2(r) \rfloor$. Then, there exist variables $u_1, \dots, u_{k-1} \geq 0$ such that each constraint $z^s \geq \varepsilon^r$ in $\text{LR}_{\Phi, \|\cdot\|}$ can be equivalently written as constraints in the form:

$$\begin{aligned} u_j^2 &\leq u_l^{a_j} z^{b_j} \varepsilon^{c_j}, \\ \varepsilon^2 &\leq u_h u_{h-1}^{d_h} z^{f_h} \varepsilon^{g_h}, \\ u_j &\geq 0 \end{aligned}$$

with $j = 1, \dots, k-1$ and such that $1 \leq a_j + b_j + c_j \leq 2$ for given $a_j, b_j, c_j \in \mathbb{Z}_+$ and $d_h, f_h, g_h \in \mathbb{Z}_+$ such that $d_h + b_h + c_h = 1$.

By the above lemma, the nonlinear constraints in the form $z^s \geq \varepsilon^r$ are written as second order cone constraints in the form $X^2 \leq YZ$ or $X^2 \leq Y$ (for some choices of the variables X, Y and Z in our model). These constraints are then equivalent to one of the following two semidefinite constraints:

$$\begin{pmatrix} Y+Z & 0 & 2X \\ 0 & Y+Z & Y-Z \\ 2X & Y-Z & Y+Z \end{pmatrix} \succeq 0, Y+Z \geq 0 \text{ or } \begin{pmatrix} Y & 0 & 2X \\ 0 & Y & Y \\ 2X & Y & Y \end{pmatrix} \succeq 0, Y \geq 0.$$

Hence, the difficulty of solving Problem $\text{LR}_{\Phi, \|\cdot\|}$, depends essentially on the choice of the residuals since all except constraints (6) are linear or second order cone constraints which can be efficiently handled with nowadays modern optimization techniques. In the next sections we analyze different choices of the residuals.

Remark 8 (Subset Selection and Regularization). *In the case where the number of points (n) is much smaller than the dimension of the space (d), it is common in Statistics to compute fitting hyperplanes over a smaller dimensional space. The new space is determined by those components that, after projecting, allows a good fitting when it is compared to the dimension of the new space. Several methods have been proposed in the recent literature to perform such a computation. If the dimension of the new space, $q < d$, is given, a constraint in the form $\|\beta_{-0}\|_0 \leq q$ (here $\|\cdot\|_0$ stands for the support function or nuclear norm, i.e., the number of nonzero components of the vector) may be included in the mathematical programming formulation (see [27, 8]), which gives rise to the so called Subset Selection Problem. If such a dimension is not known, regularization methods that penalize the number of nonzero elements or the size of β_{-0} can be applied to solve the Feature Selection Problem (see [29]). Note that both types of approaches can be easily incorporated in our models.*

2.1. Goodness of Fitting. After addressing the problem of locating/fitting a hyperplane with respect to a set of points, we will analyze the goodness of this fitting extending the well known coefficient of determination in Regression Analysis. For the sake of presentation, we assume that the variable that needs to be analyzed in terms of dependence to the others is the last coordinate X_d , or in other words $Y = X_d$. The *goodness of fitting index* is defined as:

$$\text{GCoD}_{\Phi, \varepsilon} = 1 - \frac{\Phi^*}{\Phi_0^*},$$

where Φ^* is the optimal value of $(\text{FHP}(\Phi, \varepsilon))$, namely $\Phi(\varepsilon_x(\hat{\beta}))$, and Φ_0^* is the optimal value of $\text{FHP}(\Phi, \varepsilon)$

when it is additionally required that β is in the form $\beta = (\beta_0, \overbrace{0, \dots, 0}^{d-1}, -1)$, i.e. the hyperplane is imposed to be constant ($X_d = \beta_0$). Note that the components $1, \dots, d-1$ do not appear in the model. Hence, Φ_0^*

measures the global error assumed by the *best* fitting “vertical” hyperplane; whereas $\text{GCoD}_{\Phi, \varepsilon}$ measures the improvement of the model that considers all the dimensions with respect to the one that omits all (except one) of them. Observe that this coefficient coincides with the classical coefficient of determination provided that the aggregation criteria is the overall sum and the residuals are the squared vertical distances: in that case $\hat{\beta}_0 = \bar{x}_{\cdot d}$ (the sample mean of the *dependent* variable).

The GCoD clearly verifies one of the important properties of the standard coefficient of determination, $0 \leq \text{GCoD}_{\Phi, \varepsilon} \leq 1$. Furthermore, one may interpret the coefficient as a measure of how good is the best possible hyperplane under certain criterion and residual choice with respect to the best *horizontal* hyperplane. When GCoD is close to 0, it is because $\Phi^* \simeq \Phi_0^*$, so not appreciable improvement is given by the complete model (which considers all the components) with respect to the simple constant model; whenever GCoD is close to 1, it means that $\Phi^* \ll \Phi_0^*$, being the proposed model significantly better than the constant model (note that $\text{GCoD} = 1$ iff $\Phi^* = 0$, i.e., when the model perfectly fits the demand points). Hence, the closer the GCoD to one, the better the fitting; whereas the closer to zero, the better is the constant model with respect to the full model.

Observe that the above definition coincides with some of the choices to measure the goodness of fitting for robust alternatives to the least sum of squares methodology (see [28]).

To obtain the GCoD, apart from solving $\text{FHP}(\Phi, \varepsilon)$ to get Φ^* , we must also solve the problem:

$$(12) \quad \Phi_0^* = \min_{\beta_0 \in \mathbb{R}} \Phi(D(x_1, \mathcal{H}_0), \dots, D(x_n, \mathcal{H}_0)),$$

where $\mathcal{H}_0 = \{y \in \mathbb{R}^d : y_d = \beta_0\}$ for some $\beta_0 \in \mathbb{R}$.

Lemma 9. *If the residual mapping $\varepsilon_x : \mathbb{R}^{d+1} \rightarrow \mathbb{R}_+$ is induced by a norm $\|\cdot\|$. Then, Problem (12) is equivalent to*

$$(\text{LRP}_{\Phi, \varepsilon}^0) \quad \Phi_0^* = \min_{\beta_0 \in \mathbb{R}} \Phi(\kappa_\varepsilon |x_{1d} - \beta_0|, \dots, \kappa_\varepsilon |x_{nd} - \beta_0|),$$

where

$$\kappa_\varepsilon = \frac{1}{\max_{z \in \mathbb{R}^d : \|z\| \leq 1} z_d}$$

Proof. For the point x_k in the data set, the residual under the assumption $X_d = \beta_0$ is $\varepsilon_{x_k}(\beta_0) = D(x_k, \mathcal{H}_0) = \min_{y \in \mathcal{H}_0} \|x_k - y\|$, where $\mathcal{H}_0 = \{y \in \mathbb{R}^d : y_d = \beta_0\}$ for some $\beta_0 \in \mathbb{R}$. Then, by (2) in Lemma 1

$$\varepsilon_{x_k}(\beta_0) = \frac{|x_{kd} - \beta_0|}{\|(0, \dots, 0, -1)\|^*}$$

with $\|\cdot\|^*$ the dual norm of $\|\cdot\|$. By definition of the dual norm $\|y\|^* = \max_{z \in \mathbb{R}^d : \|z\| \leq 1} z^t y$. Hence, applying such a definition to $y = (0, \dots, 0, -1)$ the result follows. \square

From the above result it is easy to see that $\kappa_\varepsilon = 1$, provided that ε_x is induced by any ℓ_p norm, even for the ℓ_1 and the ℓ_∞ cases. However, as we will see in Section 4, not all the norms have the same κ_ε constant.

Next, with our specifications for Φ , given by $\text{FHP}(\Phi, \varepsilon)$, the problem to be solved to obtain Φ_0^* is:

$$(\text{LRP}_{\lambda, p}^0) \quad \Phi_0^* = \kappa_\varepsilon \min_{\beta_0 \in \mathbb{R}} f_{\lambda, p}(\beta) := \sum_{i=1}^n \lambda_i \varepsilon_i^p$$

where $\varepsilon_i = |x_{id} - \beta_0|$ for $i = 1, \dots, n$.

Solutions to Problem $\text{LRP}_{\lambda, p}^0$ for a given $\beta_0 \in \mathbb{R}$ motivate the introduction of the concept of *ordered median point*. Indeed, β_0 is a (λ, p) -ordered median point ((λ, p) -omp in short) if it is an optimal solution to $\text{LRP}_{\lambda, p}^0$.

Some special cases of (λ, p) -omp are well-known and widely used in the so-called location analysis literature. If $\lambda_i = 1$ for all $i = 1, \dots, n$, the $(\lambda, 1)$ -omp is known to coincide with the median, $\text{median}(x_{1d}, \dots, x_{nd})$, of $\{x_{1d}, \dots, x_{nd}\}$; while the $(\lambda, 2)$ -omp is the arithmetic mean of the $x_{\cdot d}$ -values.

In the general case, i.e. for arbitrary λ and p , the ordered median points do not have closed form expressions [17, 18], although they have been around in the field of Location Analysis for several years [31, 32]. Moreover, they can be obtained, as shown below, to be used in the computation of the goodness of fitting index.

In the following we show how to solve $\text{LRP}_{\lambda, p}^0$ for general choices of non-negative vectors λ and $p \in [1, +\infty)$. Without loss of generality we assume that $x_{1d} \leq x_{2d} \leq \dots \leq x_{nd}$. Let us denote further by $\alpha_{ik} := \frac{x_{id} + x_{kd}}{2}$ the solution of the equation $\varepsilon_i^p(\beta) = \varepsilon_k^p(\beta)$ for all $i < k$, $i, k = 1, \dots, n$ in the range (x_{1d}, x_{nd}) . Let \mathcal{A} be the set containing all the $x_{\cdot d}$ and α points and denote by z_k the k -th point in \mathcal{A} sorted in non-decreasing sequence. By construction, in the interval $I_k = (z_k, z_{k+1})$ all the functions $\varepsilon_i^p(\beta)$ are monotone for all $i = 1, \dots, n$.

Lemma 10. *The function $f_{\lambda, p}(\beta)$ has at most one critical point $\beta^* \in I_k$.*

Proof. For all $\beta \in I_k$, the function $f_{\lambda, p}$ is a non-negative linear combination of monotone functions. Therefore, its derivative can vanish in at most one point. \square

Let us denote by \mathcal{A}_c the set of all the critical points of the function $f_{\lambda,p}$ in the interval (x_{1d}, x_{nd}) . Observe that the cardinality of this set is $O(n^2)$.

Theorem 11. *For any non-negative vector λ and $p \in [1, \infty)$ the set $\mathcal{A} \cup \mathcal{A}_c$ always contains a (λ, p) -omp.*

The reader may observe that the implication of the above theorem is that the $\hat{\beta}_0$ value can be always obtained by a simple enumeration of the set $\mathcal{A} \cup \mathcal{A}_c$. Then, $\Phi_0^* = \kappa_\varepsilon \sum_{i=1}^n \lambda_i |x_{id} - \hat{\beta}_0|_{(i)}^p$. Thus, the complexity of computing GCoD is essentially the same as the resolution of Problem FHP(Φ, ε), which must be solved to obtain Φ^* .

3. CLASSICAL METHODS UNDER THE NEW FRAMEWORK

In this section we show how several classical models of fitting with hyperplanes can be cast into our general framework. We assume that we are given a set of points $\{x_1, \dots, x_n\} \subseteq \mathbb{R}^{d+1}$. In classical models in the literature, the residuals are defined as the vertical distance (with respect to the last coordinate) from the point to the hyperplane:

$$(13) \quad \varepsilon_x(\beta) = \left| x_d - \sum_{k=0}^{d-1} \frac{\beta_k}{\beta_d} x_k \right|.$$

Therefore, the difference between the considered models comes from the choice of the globalizing criterion Φ that aggregates the residuals. We have pointed out in the previous section that an important factor, in determining the difficulty of solving the mathematical programming problems for the fitting model, is the choice of the residual. This element influences much more the difficulty of the problem than the globalizing criterion. We shall show in this section how to handle, within this framework, the following 4 well-regarded models: Least Sum of Squares (LSS), Least Sum of Absolute Deviation (LAD), Least Quantile of Squares (LQS) and Least Trimmed Sum of Squares (LTS). These four well-known models are presented below as particular cases of our general framework described in FHP(Φ, ε).

A particularity of the models where the residuals are measured as the vertical distance between the point and the hyperplane, is that the response for a given data z coincides with $\hat{z} = z_d - \sum_{k=0}^{d-1} \frac{\hat{\beta}_k}{\hat{\beta}_d} z_k$, which is the direct evaluation of z over the linear function that defines the fitted hyperplane. This property will not be valid, in general, for residuals different from the vertical distance.

3.1. Least Sum of Squares fitting problem. We start our analysis with the LSS method, credited to Gauss. It is the most widely used approach to estimate the coefficients of a linear model because its simplicity and its theoretical implications for the inference over the total population. However, somehow restricting hypotheses are required in order to be applied (see e.g. [19]).

The LSS criterion is defined as the sum of the squares of the residuals, that is:

$$\Phi_{LSS}(\varepsilon_1, \dots, \varepsilon_n) = \sum_{i=1}^n \varepsilon_i^2,$$

where the residuals ε_k are given by (13).

In case $n > d$, assuming without loss of generality that $\beta_d = 1$, and that the given points are linearly independent, the optimality conditions of the problem allow to compute the best LSS parameters as:

$$\beta = (X^t X)^{-1} X^t y$$

where X is the $n \times d$ -matrix obtained from the sample data by columns and $y^t = (x_{1d}, \dots, x_{nd}) \in \mathbb{R}^n$ are the responses of the last component of the model. Hence, the complexity of computing the parameters under the LSS method is $O(nd^2)$ which results from the complexity of multiplying $n \times d$ matrices. However, even though there is a closed form formula, it may appear numerical errors when computing the inverse of the matrix $X^t X$ if the rows of X are linearly dependent or close to the linear dependence. Alternatively, one can compute the parameter β , regardless of the degree of dependence of the variables in the model by solving either a quadratic programming or a second order cone programming problem; which is nowadays doable with on-the-shell software.

Theorem 12. *An optimal parameter $\beta^* \in \mathbb{R}^d$ that minimizes Φ_{LSS} can be obtained by solving any of the following two problems:*

$$\begin{array}{lcl} \text{(LSS}_{QP}\text{)} & \min \sum_{i=1}^n z_i^2, & \\ & z_i \geq x_{id} - \sum_{k=1}^{d-1} \beta_k x_{ik} - \beta_0, & \\ & \beta \in \mathbb{R}^d, z \in \mathbb{R}_+^n. & \end{array} \quad \left| \begin{array}{l} \text{(LSS}_{SOCP}\text{)} \\ (14) \end{array} \right.$$

Proof. Denote by $z_i = x_{id} - \beta_0 - \sum_{k=1}^{d-1} \beta_k x_{ik}$ and by $w_i = \left(x_{id} - \sum_{k=1}^{d-1} \beta_k x_{ik} - \beta_0 \right)^2$, for $i = 1, \dots, n$. Note that the objective functions in (LSS_{QP}) and (LSS_{SOCP}) coincide:

$$\sum_{i=1}^n z_i^2 = \sum_{i=1}^n \left(x_{id} - \sum_{k=1}^{d-1} \beta_k x_{ik} - \beta_0 \right)^2 = \sum_{i=1}^n w_i$$

Next, the minimization character of the objective function allows us to relax the equality constraint definition of the auxiliary variables to \geq -constraints and then the result follows. \square

The reader may observe that LSS corresponds to FHP(Φ, ϵ) with $\lambda^t = (1, \dots, 1)$, $p = 2$ and ϵ the vertical distance.

3.2. Least Absolute Deviation fitting problem. Another well-explored choice of residuals and criterion is the so called LAD method, introduced by Edgeworth in 1887. The globalizing criterion is the sum of the absolute value of the vertical residuals:

$$\Phi_{LAD}(\epsilon_1, \dots, \epsilon_n) = \sum_{i=1}^n |\epsilon_i|.$$

Note that LAD corresponds to the model FHP(Φ, ϵ) for with $\lambda^t = (1, \dots, 1)$ and $p = 1$. The optimal coefficients obtained with this method are known to be more robust than those by the LSS method. It follows that the mathematical programming model to be solved under this choice is:

$$(15) \quad \min_{\beta \in \mathbb{R}^{d+1}} \sum_{i=1}^n \left| x_{id} - \beta_0 - \sum_{k=1}^{d-1} \beta_k x_{ik} \right|$$

(assuming w.l.o.g. that $\beta_d = 1$).

Observe that the above problem to compute the best LAD hyperplane can be actually formulated as a linear programming problem by replacing in (LSS_{SOCP}) the quadratic constraints by those which model the absolute value.

3.3. Least Quantile of Squares fitting problem. Next, we describe another method known as Least Quantile of Squares, recently introduced by Bertsimas and Mazumder [7], which is a generalization of the Least Median of Squares (LMS) introduced by Hampel (1975). It also considers vertical distances as residuals, but the residuals are aggregated to minimize the r -quantile of the distribution of residuals (r can range in $\{1, \dots, n\}$).

$$\Phi_{LQS}(\epsilon_1, \dots, \epsilon_n) = r - \text{quantile}(\epsilon_1^2, \dots, \epsilon_n^2) := \epsilon_{(r)}^2.$$

which also fits to the general form of the aggregating criteria considered in this paper. In this case, following the

notation introduced in (1), the LQS hyperplane can be obtained for $p = 2$ and $\lambda = \left(\overbrace{0, \dots, 0}^{(r-1)}, 1, \overbrace{0, \dots, 0}^{(n-r)} \right)$. (Observe $\lfloor \frac{n}{2} \rfloor$ $\lfloor \frac{n}{2} \rfloor$)

that LMS hyperplane is also obtained within the same scheme when $p = 2$ and $\lambda = \left(\overbrace{0, \dots, 0}^{\lfloor \frac{n}{2} \rfloor}, 1, \overbrace{0, \dots, 0}^{\lfloor \frac{n}{2} \rfloor} \right)$.

Theorem 13. *An optimal parameter $\beta^* \in \mathbb{R}^{d+1}$ for LMS method can be obtained by solving the following problem:*

$$\begin{aligned} (\text{LMS}_{\text{IP}}) \quad & \min \quad \theta_r \\ & s.t. \quad (7), (9) - (11), (14), \\ & \quad \beta \in \mathbb{R}^d, z, \theta \in \mathbb{R}_+^n, w_{ij} \in \{0, 1\}, \forall i, j = 1, \dots, n, \end{aligned}$$

3.4. Least Trimmed Sum of Squares fitting problem. Finally, we present analogous formulations for the LTS method. This method was introduced by Rousseeuw [36] as a very robust alternative to the LSS method, in that it has a high breakdown point. With our notation, the residuals are again considered as the vertical distance, $p = 2$ but the aggregation criterion is now:

$$\Phi_{LTS}(\epsilon_1, \dots, \epsilon_n) = \sum_{i=1}^h \epsilon_{(i)}^2$$

where $\epsilon_{(i)} \in \{\epsilon_1, \dots, \epsilon_n\}$ with $\epsilon_{(i)} \leq \epsilon_{(i+1)}$ for $i = 1, \dots, n-1$, and $h \in \{1, \dots, n\}$. Note that in this problem one tries to minimize the sum of the h smallest squared residuals, discarding the remaining, and then, adjusting the model to the h closest points. The most common choice for h is $\lfloor \frac{n}{2} \rfloor$, considering the best 50% square residuals to compute the hyperplane (thus, discarding the other 50% of the data). The choice of h allows to control which part of the data set are *sacrificed* to find a better hyperplane. We denote by $LTS(\alpha)$ the LTS

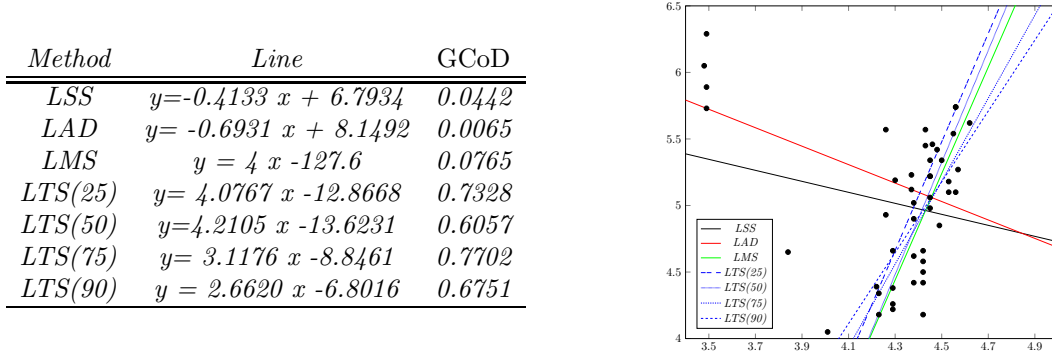


FIGURE 1. Optimal Lines with the classical methods for the stars data set.

method when $100 - \alpha\%$ of the data is discarded, i.e., the percentage of the data that may be considered as outliers.

A suitable mathematical programming formulation for the $LTS(\alpha)$ method is stated in the following result.

Theorem 14. *An optimal parameter $\beta^* \in \mathbb{R}^{d+1}$ for $LTS(\alpha)$ method can be obtained by solving the following problem:*

$$\begin{aligned}
 (LTS(\alpha)_{IP}) \quad & \min \sum_{i=1}^{\lceil \alpha n \rceil} \theta_j \\
 \text{s.t.} \quad & (7), (9) - (11), (14), \\
 & \beta \in \mathbb{R}^{d+1}, z, \theta \in \mathbb{R}^n, w_{ij} \in \{0, 1\}, \forall i, j = 1, \dots, n.
 \end{aligned}$$

We illustrate the differences of the above classical models in a well-known data set that appears in [37]. The algorithms were implemented in R with the Gurobi callable library.

Example 15. *The data considered in this example consists of 47 points in \mathbb{R}^2 about stars of the CYG OB1 cluster in the direction of Cygnus [42]. The first coordinate, X_1 , is the logarithm of the effective temperature at the surface of the star and the second one, X_2 , is the logarithm of its light intensity. This data set has also been analyzed in [37] and [48], among others.*

We run the LSS, LAD, LMS and $LTS(\alpha)$ with $\alpha \in \{25, 50, 75, 90\}$. The obtained lines and the goodness of fitting ($GCoD_{\Phi, \epsilon}$) are shown in Figure 1.

Observe that the LSS and LAD models were not able to adequately fit the data while the others (which are somehow similar) show their better performance against the outliers. Note also that GCoD reflects this fact, although it is not clear whether LTS(75) (the one with the largest GCoD) is better than the others.

In order to show the behavior of the LTS models and which are the results of their optimal fitting lines, Figure 1 shows the fitting lines that minimize the 25%, 50%, 75% or 90% of the residuals and the points that the corresponding optimization problems discard (filled dots in the subfigures) to reach the fitted lines.

Apart from the classical models described above, the standard vertical distance residuals may be aggregated using a general Φ function as those introduced in (1) providing a wide family of new methods to compute the coefficient of the best fitting hyperplanes. Also, linear constrained versions of the above methods may be considered by adding the adequate constraints to the corresponding formulations. Furthermore, many other alternative methods that use vertical distance residuals as MINSADBED or convex combinations of LSS and LAD methods [2] can easily be cast into our modelling framework. The formulations that allow solving those problems are rather similar to those already presented in this section and therefore are left for the interested reader.

4. FITTING HYPERPLANES WITH BLOCK-NORM RESIDUALS

In this section we present models to compute the parameters of fitting hyperplane when the distance point-to-hyperplane is assumed to be a block-norm distance between the point and the closest point in the hyperplane; and the aggregation criterion is considered in the general form given by $FHP(\Phi, \epsilon)$. Recall that a block norm is a norm such that its unit ball is a polytope symmetric with respect to the origin and with non empty interior. Block norms, also referred to as polyhedral norms, play an important role in the measurement of distances in many areas of Operations Research and Applied Mathematics as for instance in Location analysis or Logistics. They are often used to model real world situations (like measuring highway distances) more accurately than the standard Euclidean norm. In addition, they can also be used to approximate arbitrary norms since the set of block norms is dense in the set of all norms [47].

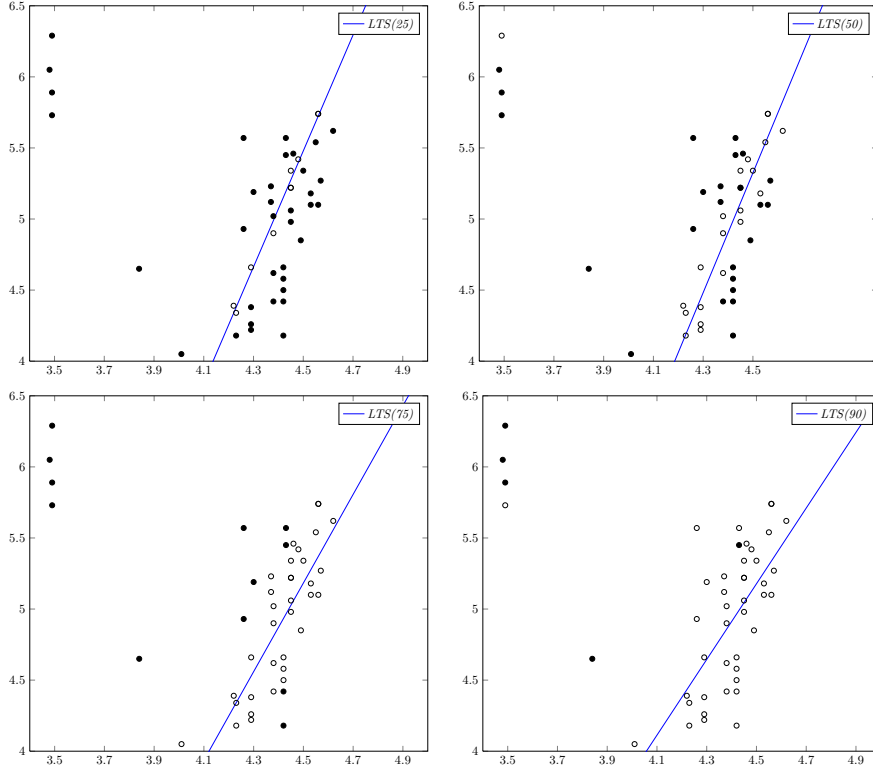


FIGURE 2. Estimated models and discarded points in LTS models.

We denote by $\|\cdot\|_B$ the norm in \mathbb{R}^d whose unit ball is given by a symmetric with respect to the origin, with non empty interior polytope B , i.e. $B = \{x \in \mathbb{R}^d : \|x\|_B \leq 1\}$. Let $\text{Ext}(B) = \{b_g : g = 1, \dots, G\}$ be the set of extreme points of B and B^0 the polar set of B which is defined as:

$$B^0 = \{v \in \mathbb{R}^d : v^t b_g \leq 1, g = 1, \dots, G\}$$

and $\text{Ext}(B^0) = \{b_1^0, \dots, b_{G^0}^0\}$.

The following result characterizes the expression of a block-norm distance in terms of the extreme points of the polar set of the polytope B .

Lemma 16 (Ward and Wendell [46, 47]). *Let B be a polytope in \mathbb{R}^d and $x \in \mathbb{R}^d$, then:*

$$\|x\|_B = \max\{|x^t b_g^0| : g = 1, \dots, G^0\}.$$

Special cases of block norms are the Manhattan (ℓ_1) and the Chebyshev (ℓ_∞) norms for adequate choices of the extreme points of the unit balls. For instance in \mathbb{R}^2 , such distances are characterized by the following set of extreme points of their unit balls, $\{\pm(1, 0), \pm(0, 1)\}$ and $\{\pm(1, 1), \pm(1, -1)\}$, respectively. Any block norm $\|\cdot\|_B$ in \mathbb{R}^d induces a distance between vectors $x, y \in \mathbb{R}^d$ given by $D_B(x, y) = \|x - y\|_B$.

Given a set of points $\{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ and a polyhedral unit ball B , our goal is to obtain the hyperplane $\mathcal{H}(\beta) = \{y \in \mathbb{R}^d : (1, y^t)\beta = 0\}$ such that the overall distance $D_B(\cdot, \cdot)$ from the sample to $\mathcal{H}(\beta)$ is minimized according to the globalizing criterion Φ (for $1 \leq p = \frac{r}{s} \in \mathbb{Q}$). That is:

$$(\text{RM}_B) \quad \min_{\beta \in \mathbb{R}^{d+1}} \sum_{i=1}^n \lambda_i \varepsilon_{(i)}^p$$

where for any $x \in \mathbb{R}^d$, $\varepsilon_x = D_B(x, \mathcal{H}(\beta)) = \min_{z \in \mathcal{H}(\beta)} D_B(x, z)$, is the “ $\|\cdot\|_B$ -projection” of x onto the hyperplane $\mathcal{H}(\beta)$, and $\varepsilon_{(i)}$ denotes the element in $\{\varepsilon_{x_1}, \dots, \varepsilon_{x_n}\}$ which is sorted in the i -th position (in nondecreasing order).

We recall that according to equation (2) in Lemma 1, for any polytope B symmetric with respect to the origin and with non empty interior, and $\mathcal{H}(\beta) = \{y^t \in \mathbb{R}^d : (1, y^t)\beta = 0\}$ then $D_B(x_{-0}, \mathcal{H}(\beta)) = \frac{|\beta^t x|}{\|\beta_{-0}\|_{B^0}}$, where B^0 is the polar set of B and $x^t = (1, x_1, \dots, x_d) \in \mathbb{R}^{d+1}$ is a given point.

Lemma 17. *Let $\beta^* \in \mathbb{R}^{d+1}$ be an optimal solution of RM_B with $\beta_{-0}^* \neq 0$. Then, $\beta' = \frac{\beta^*}{\|\beta_{-0}^*\|_{B^0}}$ is also an optimal solution of RM_B with $\|\beta'_{-0}\|_{B^0} = 1$. Thus, there is an optimal solution of RM_B , β , that verifies $D_B(x_{-0}, \mathcal{H}(\beta)) = |\beta^t x|$ for any $x^t = (1, x_1, \dots, x_d) \in \mathbb{R}^{d+1}$.*

From the above lemma, we have

Theorem 18. *Let $\{x_1, \dots, x_n\} \subset \mathbb{R}^{d+1}$ be a set of points and let $B \subset \mathbb{R}^d$ be a polytope with $\text{Ext}(B) = \{b_1, \dots, b_G\}$. Then, RM_B is equivalent to the following disjunctive programming problem*

$$\begin{aligned}
 (\text{LRP}_{\Phi, B}) \quad & \rho^*(B) := \min \sum_{j=1}^n \lambda_j \theta_j \\
 \text{s.t.} \quad & (7) - (11) \\
 (16) \quad & \varepsilon_i \geq \beta^t x_i, \forall i = 1, \dots, n, \\
 (17) \quad & \varepsilon_i \geq -\beta^t x_i, \forall i = 1, \dots, n, \\
 (18) \quad & \beta_{-0}^t b_g \leq 1, \forall g = 1, \dots, G, \\
 (19) \quad & \bigvee_{g=1}^G \beta_{-0}^t b_g = 1, \\
 & \beta \in \mathbb{R}^{d+1}, z, \theta, e \in \mathbb{R}^n. \\
 & w_{ij} \in \{0, 1\}, \forall i, j = 1, \dots, n.
 \end{aligned}$$

Proof. Let us denote by $\varepsilon_i = D_B(x_i, \mathcal{H}(\beta))$. By Lemma 1, $\varepsilon_i = \frac{|\beta^t x_i|}{\|\beta_{-0}\|_{B^0}}$. Furthermore, by Lemma 17, we can assume that $\|\beta_{-0}\|_{B^0} = 1$, hence $\varepsilon_i = |\beta^t x_i|$ (constraints (16) and (17)). By Lemma 16, $\|\beta_{-0}\|_{B^0} = \max\{|\sum_{i=1}^d \beta_i b_{gi}| : g = 1, \dots, G\}$ since $(B^0)^0 = B$. Hence, there exists $g_0 \in \{1, \dots, G\}$ such that $\|\beta_{-0}\|_{B^0} = 1$ (disjunctive constraint (19)) and thus $\sum_{k=1}^d \beta_k b_{g_0 k} \leq \sum_{k=1}^d \beta_k b_{g_0 k} = 1$ (constraint (18)). (Note that absolute values do not need to be taken explicitly into account since if $b_g \in \text{Ext}(B)$, then $-b_g \in \text{Ext}(B)$.) \square

The above problem can be equivalently written as an unique mixed integer second order cone programming problem once constraints (8) are transformed using the result in Lemma 7 and binary variables are added to decide which g_0 is chosen to verify constraint (18). By the same token, this problem can be also equivalently rewritten as G different SOCP programming problems (each of them fixed to verify one of the disjunctive constraints). Furthermore, MINLP disjunctive programming techniques (e.g. [4], [22]) may be used to solve the corresponding problem. The following result states a MINLP formulation for RM_B :

Corollary 19. *Let $\{x_1, \dots, x_n\} \subset \mathbb{R}^{d+1}$ be a set of points and let $B \subset \mathbb{R}^d$ be a polytope with $\text{Ext}(B) = \{b_1, \dots, b_G\}$. Then, $\text{LRP}_{\Phi, B}$ is equivalent to the following problem:*

$$\begin{aligned}
 (\text{LRP}_{\Phi, B}) \quad & \rho^*(B) := \min \sum_{j=1}^n \lambda_j \theta_j \\
 \text{s.t.} \quad & (7) - (11) \\
 (20) \quad & \varepsilon_i \geq \beta_h^t x_i, \forall i = 1, \dots, n, h = 1, \dots, G, \\
 (21) \quad & \varepsilon_i \geq -\beta_h^t x_i, \forall i = 1, \dots, n, h = 1, \dots, G, \\
 (22) \quad & \beta_{-0h}^t b_g \leq 1, \forall g = 1, \dots, G, h = 1, \dots, G, \\
 (23) \quad & \beta_{-0h}^t b_h = \xi_h, h = 1, \dots, G, \\
 (24) \quad & \sum_{h=1}^G \xi_h = 1, \\
 & z, \theta, \varepsilon \in \mathbb{R}^n, \\
 & \beta_h \in \mathbb{R}^{d+1}, \xi_h \in \{0, 1\}, \forall h = 1, \dots, G, \\
 & w_{ij} \in \{0, 1\}, \forall i, j = 1, \dots, n.
 \end{aligned}$$

Some special cases for the globalizing criterion Φ allow even simpler formulations reducing considerably the computational complexity of the problems. In particular, when $\lambda_i = 1$ for all $i = 1, \dots, n$, the integer variables representing ordering (w_{ij}) can be removed from the above formulation.

The following result will allow us to consider polyhedral norms which are *dilations* of other polyhedral norms, i.e., polyhedral norms $\|\cdot\|_{\mu B}$ for some bounded polyhedron B and $\mu > 0$ ($\mu B = \{\mu z : z \in B\}$).

Corollary 20. *Let \overline{B} be a polytope and $\mu > 0$. Then, if β^* is an optimal solution for $\text{LRP}_{\Phi, B}$ for $B = \overline{B}$, $\hat{\beta} = \frac{1}{\mu} \beta^*$ is an optimal solution for $\text{LRP}_{\Phi, B}$ when $B = \mu \overline{B}$. Moreover, $\rho^*(\mu \overline{B}) = \frac{1}{\mu^p} \rho^*(\overline{B})$.*

Proof. It is sufficient to observe that for any $\beta \in \mathbb{R}^{d+1}$:

$$\begin{aligned} \|(\beta_1, \dots, \beta_d)\|_{\mu\bar{B}} &= \max\{|\mu b_g^t \beta^t| : g = 1, \dots, G\} \\ &= \mu \max\{|b_g^t \beta^t| : g = 1, \dots, G\} = \mu \|(\beta_1, \dots, \beta_d)\|_{\bar{B}^0}. \end{aligned}$$

Since $\Phi_{\mu\bar{B}}(\varepsilon_1, \dots, \varepsilon_n) = \frac{1}{\mu^p} \Phi_{\bar{B}}(\varepsilon_1, \dots, \varepsilon_n)$, we get the relation between the optimal values. Let β^* be an optimal solution of $\text{LRP}_{\Phi, B}$. Then, $\frac{1}{\mu} \beta^*$ is clearly a feasible solution to $\text{LRP}_{\Phi, \bar{B}}$ when $B = \mu\bar{B}$ since $\|(\frac{1}{\mu} \beta_1^*, \dots, \frac{1}{\mu} \beta_d^*)\|_{\mu\bar{B}^0} = \|(\beta_1^*, \dots, \beta_d^*)\|_{\bar{B}^0} = 1$. \square

For the sake of computing GCoD, for solutions to problems with block-norm residuals, note that the one dimensional problem $\text{LRP}_{\Phi, \varepsilon}^0$ does depend on Φ and also on the residuals through κ_ε . Let us denote by κ_B the constant κ_ε when the residuals ε_x are defined as the block-norm projection with unit ball given by the polytope B .

Corollary 21. *Let $B \subset \mathbb{R}^d$ be a polytope. The Goodness of fitting index, GCoD, when the residuals are defined as the block-norm distance with unit ball B , can be computed as:*

$$\text{GCoD}_{\Phi, \varepsilon} = 1 - \frac{\Phi^*}{\sum_{i=1}^n |x_{id} - ((\lambda, p) - \text{omp}(x_{\cdot d}))|^p} \cdot \max_{g=1, \dots, G} |b_{gd}|,$$

where $(\lambda, p) - \text{omp}(x_{\cdot d})$ is the solution to the problem $\text{LRP}_{\Phi, \varepsilon}^0$ with residuals measured with the polyhedral norm with unit ball B .

Proof. By Lemma 9 the goodness of fitting index $\text{GCoD}_{\Phi, \varepsilon}$ can be computed as:

$$(25) \quad \text{GCoD}_{\Phi, \varepsilon} = 1 - \frac{\Phi^*}{\min_{\beta_0 \in \mathbb{R}} \Phi(\kappa_B |x_{1d} - \beta_0|, \dots, \kappa_B |x_{nd} - \beta_0|)},$$

where $\kappa_B = \frac{1}{\max_{z \in B} z_d}$.

Observe that since B is a polytope then the above maximum is attained in an extreme point of B , namely b_1, \dots, b_G ; and thus $\kappa_B = \frac{1}{\max_{g=1, \dots, G} b_{gd}}$.

Next, the problem $\text{LRP}_{\Phi, \varepsilon}^0$ in this case can be expressed as:

$$\kappa_B \min_{\beta_0 \in \mathbb{R}} \sum_{i=1}^n \lambda_i |x_{id} - \beta_0|_{(i)}^p.$$

Recall that this is a (λ, p) Ordered median problem and that its optimal solution, a (λ, p) -omp, can be easily obtained by the result in Theorem 11. Replacing the optimal solution to this problem in (25) it results in:

$$\text{GCoD}_{\Phi, \varepsilon} = 1 - \frac{\Phi^*}{\sum_{i=1}^n |x_{id} - ((\lambda, p) - \text{omp}(x_{\cdot d}))|^p} \cdot \max_{g=1, \dots, G} |b_{gd}|.$$

\square

Note that for $\lambda = (1, \dots, 1)$ the $(\lambda, 1)$ -omp is the standard median point and thus the expression $\sum_{i=1}^n |x_{id} - \text{median}(x_{\cdot d})|$ is what it is usually called the *mean absolute deviation with respect to the median*. It is a well-known criterion to find robust optimal hyperplanes of the mean value and a direct measure of the scale of a random variable about its median with many applications in different fields (see [33]).

We illustrate the behavior of the block-norm residuals fitting hyperplanes with the same data set used in the Section 3.

Example 22. *We consider again the stars data used in Example 15. In this case, we run our implementation in R for ℓ_1 -norm, ℓ_∞ -norm and hexagonal norm (as the one used in [32] with $\text{Ext}(B) = \{\pm(2, 0), \pm(2, 2), \pm(-1, 2)\}$) residuals. We use three different criteria: overall SUM ($\lambda = (1, \dots, 1)$ and $p = 1$), MAXimum ($\lambda = (1, 0, \dots, 0)$*

and $p = 1$), K -centrum ($\lambda = (\overbrace{0, \dots, 0}^K, \overbrace{1, \dots, 1}^{n-K})$) for $K = \lfloor 0.75n \rfloor$ (the model will minimize the sum of the 25% greatest residuals) and anti- K -centrum ($\lambda = (\overbrace{1, \dots, 1}^K, \overbrace{0, \dots, 0}^{n-K})$) for $K = \lfloor 0.5n \rfloor$ (the model will minimize the sum of the 50% smallest residuals). The results for all the combinations and the graph for the K -centrum lines are shown in Figure 3.

Method (Φ, ε)	Optimal Line	GCoD $_{\Phi, \varepsilon}$
(SUM, ℓ_1)	$y = 7x - 25.81$	0.6505853
(SUM, ℓ_∞)	$y = 5.25x + -18.1425$	0.7009688
(SUM, Hex)	$y = 7x - 25.81$	0.6505853
(MAX, ℓ_1)	$y = -3.230769x + 18.77577$	0.5336373
(MAX, ℓ_∞)	$y = -3.230769x + 18.77577$	0.6438685
(MAX, Hex)	$y = -3.230769x + 18.77577$	0.6438685
(kC, ℓ_1)	$y = -4.307692x + 23.03346$	0.4628481
(kC, ℓ_∞)	$y = -2.493333x + 15.67113$	0.5921635
(kC, Hex)	$y = 7.642857x + -28.67929$	0.8317972
(AkC, ℓ_1)	$y = 5.6x - 19.804$	0.8443055
(AkC, ℓ_∞)	$y = 4.869565x - 16.41565$	0.8426523
(AkC, Hex)	$y = 5.473684x - 19.28316$	0.6431602

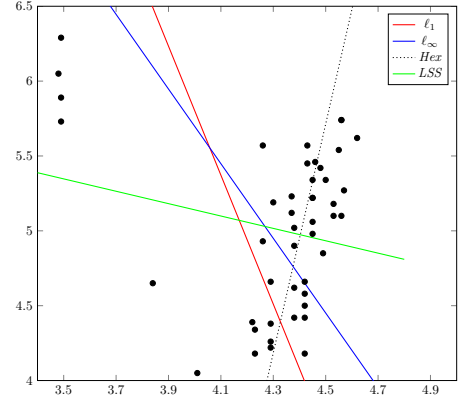


FIGURE 3. Optimal lines obtained with block-norm residuals for the stars data set.

Note that different situations may happen when running the different models: in the case of the SUM criterion the models for ℓ_1 and hexagonal residuals coincide; in the MAX criterion the three optimal lines are the same, and for the K-centrum and anti-K-centrum the three models are different. Furthermore, even in the case when the models coincide, one may have different goodness of fitting indices due to the different way of measuring distances (see the ℓ_1 and hexagonal residuals for the MAX criterion).

From the above, we observed that the GCoD are not comparable when different residuals are used in the models since the value given to the residuals (both with respect to the best model and with respect to the simplified model with only intercept) is different. Thus, the generalized coefficient allows us to compare the goodness of fitting between models provided that the distance (to measure the residuals) and the aggregation criterion are fixed.

5. FITTING HYPERPLANES WITH ℓ_τ DISTANCES

In this section we present the mathematical programming formulations for computing the optimal hyperplanes when the residuals are defined as ℓ_τ distances between demand points and the linear body. Recall that the ℓ_τ -norm in \mathbb{R}^d , with $\tau \geq 1$, is defined as:

$$\|z\|_\tau = \begin{cases} \left(\sum_{k=1}^d |z_k|^\tau \right)^{\frac{1}{\tau}} & \text{if } \tau < \infty, \\ \max_{k=1, \dots, d} \{|z_k|\} & \text{if } \tau = \infty \end{cases}$$

for any $z = (z_1, \dots, z_d)^t \in \mathbb{R}^d$. From this norm we denote by $D_{\ell_\tau}(z, y) = \|z - y\|_\tau$ the ℓ_τ -distance between the points $z, y \in \mathbb{R}^d$. The well-known Euclidean distance that measures the straight line distance between points in \mathbb{R}^d is the ℓ_2 -norm in this family. Note that the extreme cases of ℓ_1 and ℓ_∞ represent both block and ℓ_τ -norms, since their unit balls are polytopes but also fit within the family of ℓ_τ -norms.

We recall that according to equation (2) in Lemma 1, for any $\tau = \frac{r}{s} \in \mathbb{Q}$ with $r \geq s \in \mathbb{Z}_+$, $\gcd(r, s) = 1$ and $\mathcal{H}(\beta) = \{y^t \in \mathbb{R}^d : (1, y^t)\beta = 0\}$ then $D_\tau(z, \mathcal{H}(\beta)) = \frac{|\beta^t z|}{\|\beta_{-0}\|_\nu}$ where ν is such that $\frac{1}{\tau} + \frac{1}{\nu} = 1$ (for $\tau = 1$, $\nu = \infty$ while for $\tau = \infty$, $\nu = 1$).

In this section we will assume that the residuals are defined as the shortest distance from the points to the fitted hyperplane, namely to their projections, under a given ℓ_τ norm. In other words, for a given point $\hat{x} = (1, \hat{x}_1, \dots, \hat{x}_d)^t$ the residual is: $\varepsilon_{\hat{x}}(\beta) = D_\tau(\hat{x}_{-0}, \mathcal{H}(\beta))$.

As in previous sections, for a given set of points $\{x_1, \dots, x_n\} \subseteq \mathbb{R}^{d+1}$, the computation of the parameters $\beta \in \mathbb{R}^{d+1}$ assuming that the globalizing criterion is Φ and the residuals are measured with ℓ_τ -distance can be obtained by solving an adequate optimization problem.

Theorem 23. Let $\{x_1, \dots, x_n\} \subset \mathbb{R}^{d+1}$ be a set of points, $\lambda \in \mathbb{R}^n$, $\tau = \frac{r}{s} \in \mathbb{Q}$ with $r > s \in \mathbb{N}$ and $\gcd(r, s) = 1$, and $\|\cdot\|_\tau$ a ℓ_τ -norm in \mathbb{R}^d . The Problem FHP (Φ, ε) is equivalent to the following mathematical programming problem:

$$\begin{aligned}
(\text{LPR}_{\Phi, \ell_\tau}) \quad & \min \sum_{j=1}^n \lambda_j \theta_j \\
\text{s.t.} \quad & (7) - (11), (16) - (17), \\
(26) \quad & \|\beta_{-0}\|_\nu = 1, \\
& w_{ij} \in \{0, 1\}, \quad \forall i = 1, \dots, n, \\
& z, \theta \in \mathbb{R}_+^n, \beta \in \mathbb{R}^{d+1}.
\end{aligned}$$

Note that the above problem is nonconvex for $1 < \tau < \infty$ because of the binary variables and constraint (26). Approximation schemes are available in different free and commercial solvers, although no guarantee of optimality is provided (e.g., NLOPT, MATLAB, Minotaur, ...). In what follows we describe an approximation approach based on some linear approximations of the problem.

Let P be a polyhedron such that $P \subset \mathcal{B} = \{z \in \mathbb{R}^d : \|z\|_\nu \leq 1\}$, and denote by $r_P = \sup_{\|z\|_P=1} \|z\|_\nu$ (note that by construction $r_P \leq 1$). Observe that r_P is the radius of the smallest ℓ_ν -ball containing P . In addition, let Q be a polyhedron such that $\mathcal{B} \subset Q$, and denote by $R_Q = \inf_{\|z\|_Q=1} \|z\|_\nu$ (note that by construction $R_Q \geq 1$). In this case R_Q is the radius of the largest ℓ_ν -ball contained in Q .

Theorem 24. Let $\lambda_1, \dots, \lambda_n \geq 0$ and the globalizing function $\Phi(\varepsilon_1, \dots, \varepsilon_n) = \sum_{i=1}^n \lambda_i \varepsilon_{(i)}^\delta$ then:

$$(27) \quad \Phi_{P^*} \leq \Phi_{\ell_\tau} \leq \frac{1}{r_P^\delta} \Phi_{P^*}$$

$$(28) \quad \frac{1}{R_Q^\delta} \Phi_{Q^*} \leq \Phi_{\ell_\tau} \leq \Phi_{Q^*}$$

Proof. By the relations between the norms, it is clear that $\|z\|_P \geq \|z\|_\nu \geq r_P \|z\|_P$. Let $\mathcal{H}(\beta) = \{z \in \mathbb{R}^d : (1, z^t)\beta = 0\}$. Then, for any $x \in \mathbb{R}^d$, the above relationships imply the following inequalities relating the distances with respect to $\|\cdot\|_{P^*}$ -residuals and $\|\cdot\|_\tau$ -residuals:

$$D_{P^*}(x_{-0}, \mathcal{H}(\beta)) = \frac{|\beta^t x|}{\|\beta_{-0}\|_P} \leq \frac{|\beta^t x|}{\|\beta_{-0}\|_\nu} \leq D_\tau(x_{-0}, \mathcal{H}(\beta))$$

and

$$D_\tau(x_{-0}, \mathcal{H}(\beta)) = \frac{|\beta^t x|}{\|\beta_{-0}\|_\nu} \leq \frac{|\beta^t x|}{r_P \|\beta_{-0}\|_P} \leq \frac{1}{r_P} D_{P^*}(x_{-0}, \mathcal{H}(\beta))$$

Let us consider the globalizing criterion $\Phi(\varepsilon_1, \dots, \varepsilon_n) = \sum_{i=1}^n \lambda_i \varepsilon_{(i)}^\delta$. Then, the evaluation of Φ with respect to the residuals computed with the polyhedral norm with unit ball P^* and the ℓ_τ -norm, namely $\varepsilon_{i, P^*} = D_{P^*}(x_{i, -0}, \mathcal{H}(\beta))$ and $\varepsilon_{i, \ell_\tau} = D_\tau(x_{i, -0}, \mathcal{H}(\beta))$ for all $i = 1, \dots, n$, satisfies:

$$\Phi(\varepsilon_{P^*}) \leq \Phi(\varepsilon_{\ell_\tau}) \leq \frac{1}{r_P^\delta} \Phi(\varepsilon_{P^*}).$$

This equation proves (27).

Next, it is clear that $\|z\|_Q \leq \|z\|_\nu \leq R_Q \|z\|_Q$. Now, using an argument similar to the one above we conclude that

$$\begin{aligned}
D_{Q^*}(x_{-0}, \mathcal{H}(\beta)) &= \frac{|\beta^t x|}{\|\beta_{-0}\|_Q} \geq \frac{|\beta^t x|}{\|\beta_{-0}\|_\nu} \geq D_\tau(x_{-0}, \mathcal{H}(\beta)) \\
&= \frac{|\beta^t x|}{\|\beta_{-0}\|_\nu} \geq \frac{|\beta^t x|}{R_Q \|\beta_{-0}\|_\nu} \geq \frac{1}{R_Q} D_{Q^*}(x_{-0}, \mathcal{H}(\beta)).
\end{aligned}$$

From these inequalities it clearly follows (28). \square

Let P_N be a symmetric with respect to the origin polytope with N vertices, $\{p_1, \dots, p_N\}$, inscribed in the ℓ_ν hypersphere $\mathcal{B} = \{z \in \mathbb{R}^d : \|z\|_\nu = 1\}$ and let r_{P_N} be the radius of the smallest ℓ_ν ball centered at the origin containing P_N . Let $R_{Q_N} = \frac{1}{r_{P_N}}$ and denote by Q_N the R_{Q_N} -dilation of P_N . By construction $P_N \subset \mathcal{B} \subset Q_N$.

Hence, for the globalizing function $\Phi(\varepsilon_1, \dots, \varepsilon_n) = \sum_{i=1}^n \lambda_i \varepsilon_{(i)}^\delta$, by the Theorem 24, we get that:

$$\max\{\Phi(\varepsilon_{P_N^*}), \frac{1}{R_{Q_N}^\delta} \Phi(\varepsilon_{Q_N^*})\} \leq \Phi(\varepsilon_{\ell_\tau}) \leq \min\{\Phi(\varepsilon_{Q_N^*}), \frac{1}{r_{P_N}^\delta} \Phi(\varepsilon_{P_N^*})\}$$

Furthermore, by Corollary 20, since Q_N is a dilation of P_N , both problems have the same optimal solutions and $\Phi(\varepsilon_{P_N^*}) = r_P^\delta \Phi(\varepsilon_{Q_N^*})$. Hence,

$$\frac{1}{r_{P_N}^\delta} \Phi(\varepsilon_{P_N^*}) \leq \Phi(\varepsilon_{\ell_\tau}) \leq \Phi(\varepsilon_{Q_N^*}).$$

It is clear from its definition that r_{P_N} gives the approximation error whenever a ℓ_ν -norm is replaced by a polyhedral norm with unit ball P_N . This measure can be explicitly computed from the set of inequalities that describe the polyhedron.

Lemma 25. *Let $P = \{z \in \mathbb{R}^d : a_i x \leq b_i, i = 1, \dots, N\}$ be a polytope, then:*

$$r_P = \max_{i=1, \dots, N} \frac{b_i}{\|a_i\|_\tau}.$$

Proof. First, note that $r_P = \sup_{\|z\|_P=1} \|z\|_\nu = \max_{\|z\|_P=1} \|z\|_\nu$ by the compactness of P . Thus, r_P is the ℓ_ν -inradius of P . Next, by [24], the radius of a ℓ_ν ball centered at the origin and reaching the facet $\{x \in \mathbb{R}^d : a_i^t x \leq b\}$ of P is the ℓ_ν projection of the origin onto that facet, namely $\frac{|b_i|}{\|a_i\|_\tau}$. Hence, r_P is the maximum of those distances among the N facets defining P . \square

Theorem 26. *Let $\{x_1, \dots, x_n\} \subset \mathbb{R}^{d+1}$ be a set of demand points, $\lambda \in \mathbb{R}_+^n$, $\tau = \frac{r}{s} \in \mathbb{Q}$ with $r > s \in \mathbb{N}$, $\gcd(r, s) = 1$ and the globalizing function $\Phi(\varepsilon_1, \dots, \varepsilon_n) = \sum_{i=1}^n \lambda_i \varepsilon_{(i)}^p$. The following problem provides a lower bound for Problem $\text{LPR}_{\Phi, \ell_\tau}$.*

$$\begin{aligned} (\text{Inner-}\ell_\tau) \quad & \rho^* := \min \sum_{j=1}^n \lambda_j \theta_j \\ & \text{s.t. (7) - (11)} \\ (29) \quad & \varepsilon_i \geq |\beta^t x_i|, \quad \forall i = 1, \dots, n, \\ (30) \quad & \|\beta_{-0}\|_{P_N} = 1, \\ & w_{ij} \in \{0, 1\}, \quad \forall i = 1, \dots, n, \\ & z, \theta \in \mathbb{R}_+^n, \beta \in \mathbb{R}^{d+1} \end{aligned}$$

Furthermore, $\rho^* \leq \Phi_{\ell_\tau}^* \leq \frac{1}{r_P^p} \rho^*$.

Corollary 27. *For any data set $\{x_1, \dots, x_n\} \subset \mathbb{R}^{d+1}$ and any ℓ_τ -norm with $1 < \tau < +\infty$ there exists a polyhedral norm $\|\cdot\|_B$ whose unit ball B has at most $2n$ extreme points and such that the optimal values of Problem $\text{LPR}_{\Phi, \ell_\tau}$ and $\text{LRP}_{\Phi, B}$ coincide.*

In [23] the authors propose a measure of the goodness of approximating a given norm by another norm. This measure was defined in order to quantify the approximation errors when modeling road distances between cities. We redefine this measure to evaluate the approximation errors when approximating ℓ_τ norms via polyhedral norms:

$$\text{SD} = \sum_{\substack{i=1 \\ D_\tau(x_i, \beta) > 0}}^n \frac{(D_\tau(x_i, \beta) - D_P(x_i, \beta))^2}{D_\tau(x_i, \beta)}$$

Example 28. *Let us consider again the stars data from Example 15. We run now the models using as aggregation criteria the overall sum of the residuals ($\Phi = \text{SUM}$) and the residuals are the ℓ_τ projections of the points onto the optimal line, for $\tau \in \{1.5, 2, 3\}$. The obtained estimations for the aggregation criterion $\Phi = \text{SUM}$ and their goodness of fitting ($\text{GCoD}_{\Phi, \varepsilon}$) are shown in Table 1. The obtained lines are drawn in Figure 4.*

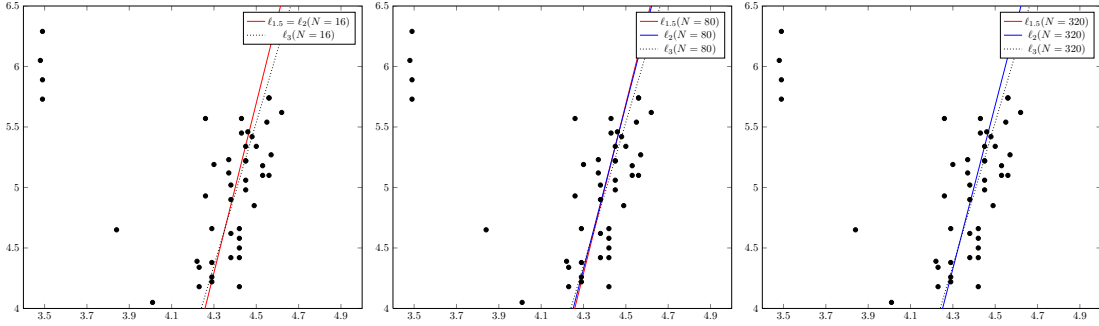
Observe that for this data set, getting high accuracy for the ℓ_τ -norm residual problems is possible using small number of vertices (N) in the approximation by polyhedral norms. As expected, increasing the number of vertices improves the accuracy at the price of increasing the computation times.

We also computed the optimal lines for different aggregation criteria ($\Phi \in \{\text{SUM}, \text{MAX}, \text{kC}, \text{AkC}\}$) with ℓ_τ residuals, $\tau \in \{1.5, 2, 3\}$, using the polyhedral approximation approach with $N = 480$ vertices. The results are shown in Table 2. The reader may observe from these results that the approximation error, although tiny, depends both of the chosen residuals and aggregation criteria.

Finally, we compare our approximation scheme for ℓ_τ residuals, on this data set, with other available implementations. Orthogonal Distance Regression (ODR) is a particular case of our general framework where ℓ_2 residuals are chosen and Φ is the sum of squares aggregation criterion (note that both approaches coincide when the coefficient of the dependent coordinate is non zero while such an assumption is not imposed in our

TABLE 1. Estimated models with minisum criterion in Example 15.

τ	N	$\hat{\beta}$	Φ^*	GCoD	R_P	r_P	Time	SD
1.5	16	(36.87, -1, 0.14)	77.1857	0.6505	0.9848	1.015	1.0	7.26×10^{-5}
	80	(36.84, -0.99, 0.14)	77.1324	0.6508263	0.9993	1.0006	1.97	6.06×10^{-6}
	320	(36.83, -0.99, 0.14)	77.1117	0.6509203	0.9999	1.0000	14.16	9.41×10^{-9}
2	16	(36.87, -1, 0.14)	77.1857	0.6505	0.9807	1.0195	1.04	7.87×10^{-3}
	80	(36.19, -0.98, 0.14)	76.3703	0.654276	0.9922	1.0007	2.01	1.91×10^{-7}
	320	(36.19, -0.98, 0.14)	76.3700	0.654277	0.9999	1.0000	16.53	1.64×10^{-7}
3	16	(34.35, -0.96, 0.16)	74.7283	0.6617	0.9801	1.0202	1.07	4.56×10^{-3}
	80	(34.09, -0.95, 0.16)	74.1627	0.66427	0.9992	1.0007	2.04	3.50×10^{-6}
	320	(34.08, -0.95, 0.16)	74.1468	0.6643	0.9999	1.0000	17.48	4.68×10^{-10}

FIGURE 4. Estimated lines for the data in Example 15 approximating by a $\{16, 80, 320\}$ -gon.

		$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	Line	$y = 5.92x - 21.1016$	$y = 6.75x - 24.6975$	$y = 7x - 25.81$
	GCoD	0.6643	0.6542	0.6509
	SD	3.36×10^{-10}	1.73×10^{-10}	1.65×10^{-9}
MAX	Model	$y = -3.2307x + 18.7757$	$y = -3.2307x + 18.7757$	$y = -3.2307x + 18.7757$
	GCoD	0.5805	0.5544	0.5381
	SD	4.07×10^{-14}	1.90×10^{-12}	3.85×10^{-13}
kC	Model	$y = -2.8133x + 16.9367$	$y = -3.1756x + 18.5100$	$y = -4.3076x + 23.0334$
	GCoD	0.5111	0.4790	0.4650
	SD	3.51×10^{-13}	7.53×10^{-10}	9.70×10^{-10}
AkC	Model	$y = 6.75x - 25.0875$	$y = 6.5555x - 24.1533$	$y = 5.175x - 17.7146$
	GCoD	0.8092	0.82512	0.8217
	SD	7.15×10^{-10}	2.10×10^{-9}	5.49×10^{-10}

TABLE 2. Optimal lines for different criteria and ℓ_τ residuals of Example 28.

models). The package `pracma` in R allows to compute ODR by using an approximated iterative procedure (see [10]). The models obtained with both approaches are shown in the following table, where one can observe that, for this data set, our approach to approximate ℓ_τ distances by polyhedral norms (with $N = 320$ vertices) has a better performance on the global error measure of the models (although as expected the models obtained by both methods are almost the same):

	ODR	SOS- ℓ_2 (SD= 9.93×10^{-11})
Model	$y = -7.05736x + 35.42935$	$y = -7.098062x + 35.60477$
Global Residuals	3.959383	3.662783

6. EXPERIMENTS

We tested the proposed models for different data sets in order to show the applicability and the differences of some of the methods detailed in the sections above. Our formulations have been coded in Gurobi 6.0 under R and executed in a PC with an Intel Core i7 processor at 2x 2.40 GHz and 4 GB of RAM. As far as we know, the battery of experiments that we performed has never been considered in the literature, since we have compared 42 different methods (several combinations of aggregation criteria and residuals measures).

6.1. Synthetic Experiments. We consider a set of randomly generated points with different peculiarities in order to test and compare the described methodologies, following similar schemes that those proposed in [7]. We generated $n = 100$ data points in dimension $d \in \{2, 4\}$, $\{x_1, \dots, x_n\} \subseteq \mathbb{R}^{d+1}$ as follows. Each x_{ik} follows an independent and identically distributed Gaussian distribution with mean 0 and standard deviation 100. We fix $\beta^t = (0, 1, \dots, 1) \in \mathbb{R}^{d+1}$. The last coordinate, x_d , is chosen as the response and we generate it as:

Aggregation criteria		Residuals
SUM	$\sum_{i=1}^n \varepsilon_i$	V
MAX	$\max_{i=1, \dots, n} \varepsilon_i$	ℓ_1
MED	$\text{median}(\varepsilon_1, \dots, \varepsilon_n)$	ℓ_∞
kC	$\sum_{i=1}^{\lfloor 0.5n \rfloor} \varepsilon_{(i)}$	$\ell_{\frac{3}{2}}$
AkC	$\sum_{i=\lfloor 0.5n \rfloor + 1}^n \varepsilon_{(i)}$	ℓ_2
SOS	$\sum_{i=1}^n \varepsilon_i^2$	ℓ_3
1.5SUM	$\sum_{i=1}^n \varepsilon_i^{\frac{3}{2}}$	

TABLE 3. Combinations of chosen aggregation criteria and residuals.

$$x_{id} = -\sum_{k=1}^{d-1} x_{ik} + u_i, \quad \forall i = 1, \dots, n,$$

where u_i is also generated as a Gaussian distribution with mean 0 and standard deviation 10.

Then, 15% of the data are now corrupted by adding an extra Gaussian term (with mean 0 and standard deviation 500) to: (1) all the components except the last one or (2) to the last coordinate.

For each one of the generated data sets, we run the models that results from the combination of the following aggregation criteria and residuals detailed in Table 3.

Tables 4-7 report, for each battery of generated data, the following information: i) the coefficients of the optimal hyperplane ($\hat{\beta}$), ii) the goodness of fitting index GCoD, iii) the percentage of the sample data which are contained in a strip delimited by two parallel hyperplanes to $y = \hat{\beta}x$ with (orthogonal) distance $\varepsilon = 10$ (%), and iv) the width of the strip that is necessary to include 90% of the data (ε_{90}).

We conclude, from the experiments for the bivariate case, that in general a better performance is observed in all the methods when the corrupted coordinate is the dependent one (Y), as compared with introducing the corruption on the independent coordinate (X). In particular, the SUM, the 1.5SUM and the kC criteria (for vertical distance residuals) get better fitting models in the Y -corrupted case. Although slightly better, almost similar results were obtained for the AkC, MEDIAN and kC (for ℓ_τ residuals) due to the robustness of those criteria. Also, we observe that for the X -corrupted case, the linear residuals (V, ℓ_1 and ℓ_∞) models coincide for all the criteria except the AkC. This is not the case in the Y -corrupted experiments, where equal or similar models were obtained for all the ℓ_τ -residuals. Observe that although in the X -corrupted case the larger % seems to imply a greater GCoD, that is not the case in the Y -corrupted experiments where one can find many combinations of criteria-residuals where that behavior does not happen.

Similar conclusions can be derived from the multivariate case ($d = 4$), except that in this case there are no coincidences between the models obtained with different combinations of criteria and residuals. Furthermore, the convenience of using measures for the goodness of fitting which are not criterion/residual dependent is confirmed.

6.2. Data: Durbin-Watson. We also performed some experiments over the classical real data sample used in [16]. The data aims to analyze the annual consumption of spirits from 1870 to 1938 ($n = 69$) from the incomes and the relative price of spirits (deflated by a cost-of-living index). Hence, the variables observed in this data sets are the logarithms (the coefficients are then interpreted in terms of percent change) of the following measures: X_1 (Real income per head), X_2 (Relative price of spirits) and X_3 (Consumption of spirits per head).

For illustrative purposes, we analyze both the global model with the three variables ($d = 3$) and the bivariate model considering X_1 and X_3 and obviating X_2 ($d = 2$).

6.2.1. Bivariate model. First, for the case $d = 2$, we run the 42 models (Table 3) over the data set where X_1 (income) and X_3 (consumption) are measured. The obtained hyperplanes are detailed in Table 8 and the fitted lines drawn in Figure 5. Note that the methods that use vertical distance residuals were not able to capture the actual behavior of the consumption with respect to the incomes. Furthermore, the MAX criterion seems to fail for any choice of residuals, since it tries to explain the unique outlier point that exists in the data set. The rest of the hyperplanes, with minimal deviations, have a similar behavior. In order to analyze the differences

TABLE 4. Results for bidimensional experiments corrupting the X variables.

		V	ℓ_1	ℓ_∞
SUM	$\hat{\beta}$	(-1.9587, 0.3011, 1)	(1.9587, -0.3011, -1)	(0.4240, -0.9403, -1)
	GCoD	0.1456	0.1456	0.5342
	%	8%	8%	65%
	ϵ_{90}	141.2995	141.2995	87.0871
MAX	$\hat{\beta}$	(10.9038, 0.1571, 1)	(10.9038, 0.1571, 1)	(10.9038, 0.1571, 1)
	GCoD	0.1484	0.1484	0.2641
	%	10%	10%	10%
	ϵ_{90}	158.9295	158.9295	158.9295
SOS	$\hat{\beta}$	(-3.1753, 0.1860, 1)	(3.1753, -0.1860, -1)	(-1.8549, 0.2858, 1)
	GCoD	0.2261	0.2261	0.4925
	%	8%	8%	9%
	ϵ_{90}	157.7177	157.7177	143.1279
1.5SUM	$\hat{\beta}$	(-3.5386, 0.2112, 1)	(3.5397, -0.2112, -1)	(0.3967, -0.4136, -1)
	GCoD	0.1812	0.1812	0.4499
	%	8%	8%	8%
	ϵ_{90}	152.361	152.3626	127.4389
kC	$\hat{\beta}$	(-3.0188, 0.2328, 1)	(-3.0188, 0.2328, 1)	(0.3503, 0.9091, 1)
	GCoD	0.1226	0.1226	0.4275
	%	8%	8%	60%
	ϵ_{90}	150.5599	150.5599	85.1974
AkC	$\hat{\beta}$	(5.8180, 0.7718, 1)	(2.2956, 0.7734, 1)	(2.6795, 0.9874, 1)
	GCoD	0.6735	0.9040	0.9758
	%	29%	34%	70%
	ϵ_{90}	77.4723	74.8420	92.8187
MED	$\hat{\beta}$	(6.1846, 0.7795, 1)	(6.1842, 0.7795, 1)	(1.3314, 0.9890, 1)
	GCoD	0.7021	0.8690	0.9741
	%	31%	31%	70%
	ϵ_{90}	78.4775	78.4772	91.9773

		$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	$\hat{\beta}$	(-0.2603, -0.9299, -1)	(-0.2603, -0.9299, -1)	(-0.2603, -0.9299, -1)
	GCoD	0.4133	0.3417	0.2615
	%	62%	62%	62%
	ϵ_{90}	86.7791	86.7791	86.7791
MAX	$\hat{\beta}$	(-10.9038, -0.1571, -1)	(-10.9038, -0.1571, -1)	(10.9038, 0.1571, 1)
	GCoD	0.1821	0.1588	0.1495
	%	10%	10%	10%
	ϵ_{90}	158.9295	158.9295	158.9295
SOS	$\hat{\beta}$	(2.4728, -0.2391, -1)	(-2.8551, 0.2102, 1)	(-3.1181, 0.1903, 1)
	GCoD	0.3163	0.2552	0.2295
	%	8%	8%	8%
	ϵ_{90}	149.8204	151.9362	156.6873
1.5SUM	$\hat{\beta}$	(3.4138, -0.2225, -1)	(3.0670, -0.2704, -1)	(1.4864, -0.3260, -1)
	GCoD	0.1853	0.2145	0.2799
	%	8%	9%	7%
	ϵ_{90}	149.6913	145.969	135.7776
kC	$\hat{\beta}$	(-2.6422, 0.2474, 1)	(-0.2632, -0.9011, -1)	(-0.3503, -0.9091, -1)
	GCoD	0.1263	0.1913	0.2791
	%	9%	57%	60%
	ϵ_{90}	147.9623	84.4867	85.1974
AkC	$\hat{\beta}$	(-0.0741, 0.9357, 1)	(2.2028, 1.0126, 1)	(-0.9506, 0.9930, 1)
	GCoD	0.9468	0.9576	0.9645
	%	64%	70%	65%
	ϵ_{90}	86.9840	94.2569	91.5147
MED	$\hat{\beta}$	(1.5779, -0.9545, -1)	(2.9207, 1.0139, 1)	(0.2899, 0.9792, 1)
	GCoD	0.9530	0.9611	0.9655
	%	63%	69%	65%
	ϵ_{90}	88.5178	94.8548	90.5271

between these models we also report in Table 9 the marginal variations of each one of the models (according to Lemma 1).

Observe that, when the ℓ_1 residuals are considered, all except the MAX criterion provide a 0 marginal variation. This pattern can be explained as a result of Lemma 2 and the fact that the ℓ_1 -norm unit ball in \mathbb{R}^2 has extreme points $\{\pm(0, 1), \pm(1, 0)\}$. Hence $k(\beta) = \begin{cases} 1 & \text{if } \beta_2 = \max\{|\beta_1|, |\beta_2|\}, \\ -1 & \text{if } \beta_2 = -\max\{|\beta_1|, |\beta_2|\}, \\ 0 & \text{otherwise.} \end{cases}$. Thus, the marginal

variation of X_1 with respect to X_3 is zero iff $|\beta_1| = \max\{|\beta_1|, |\beta_2|\}$, being then $|\beta_2| < |\beta_1|$. It means that the absolute value of the slope of the line is greater than 1, being the decreasing (or increasing) of the response consumption in terms of the incomes more than a 100%.

In order to validate and analyze the stability of the computed hyperplanes we perform a k -fold cross validation scheme [43] to the data set. Such a method consists of randomly partitioning the sample into k folds of similar size, S_1, \dots, S_k . For each $j \in \{1, \dots, k\}$, each optimal hyperplane is computed using the points in $\bigcup_{i \neq j} S_i$ and S_j is used to validate the results. In our case, we partitioned the data into $k = 7$ folds, each of them with 10 data, except one with 9 points. In Table 10 we summarize the results obtained with this experiment. We

TABLE 5. Results for bidimensional experiments corrupting the Y variables.

		V			ℓ_1			ℓ_∞		
		β	GCoD	%	β	GCoD	%	β	GCoD	%
SUM	β	(-0.4324, -1.0070, -1)			(-2.7476, -1.1156, -1)			(-0.8817, -1.0333, -1)		
	GCoD	0.5226			0.5464			0.7637		
	%	72%			57%			73%		
	ϵ_{90}	158.3495			144.4862			154.9621		
MAX	β	(164.40, 1.95, -1)			(-131.52, -7.30, -1)			(-131.52, -7.30, -1)		
	GCoD	0.0109			0.7575			0.7867		
	%	5%			6%			6%		
	ϵ_{90}	266.337			144.6019			144.6019		
SOS	β	(-19.4780, 0.9765, 1)			(24.3778, -3.9704, -1)			(-21.8989, 2.4558, 1)		
	GCoD	0.2459			0.8055			0.8896		
	%	24%			12%			14%		
	ϵ_{90}	176.2108			119.0515			108.3728		
1.5SUM	β	(2.2257, -0.9993, -1)			(8.1241, -2.8635, -1)			(4.2013, -1.5531, -1)		
	GCoD	0.3894			0.6583			0.8111		
	%	72%			15%			24%		
	ϵ_{90}	161.1331			114.1084			107.9904		
kC	β	(-0.6995, -0.9989, -1)			(4.8095, -1.6540, -1)			(-1.0107, -1.0744, -1)		
	GCoD	0.4422			0.4969			0.7265		
	%	71%			23%			67%		
	ϵ_{90}	159.1129			100.6695			150.2014		
AkC	β	(10.0084, -0.9838, -1)			(-1.3062, -1.0398, -1)			(-1.2815, -0.9942, -1)		
	GCoD	0.7526			0.9914			0.9961		
	%	53%			70%			72%		
	ϵ_{90}	168.5344			153.9189			159.2534		
MED	β	(8.6545, -0.9641, -1)			(-0.8028, -1.0379, -1)			(-4.3252, -1.0113, -1)		
	GCoD	0.8478			0.9894			0.9947		
	%	57%			73%			69%		
	ϵ_{90}	170.0131			154.4849			155.1026		

		$\ell_{1.5}$			ℓ_2			ℓ_3		
		β	GCoD	%	β	GCoD	%	β	GCoD	%
SUM	β	(-0.9890, -1.0403, -1)			(-0.9890, -1.0403, -1)			(-0.9890, -1.0403, -1)		
	GCoD	0.6250			0.6658			0.7023		
	%	70%			70%			70%		
	ϵ_{90}	154.0857			154.0857			154.0857		
MAX	β	(-131.52, -7.30, -1)			(-131.52, -7.30, -1)			(-131.52, -7.30, -1)		
	GCoD	0.7577			0.7598			0.7654		
	%	6%			6%			6%		
	ϵ_{90}	144.6019			144.6019			144.6019		
SOS	β	(24.0474, -3.7686, -1)			(23.2040, -3.2532, -1)			(22.5246, -2.8381, -1)		
	GCoD	0.8077			0.8195			0.8412		
	%	13%			13%			13%		
	ϵ_{90}	118.4519			119.827			115.0321		
1.5SUM	β	(8.2797, -2.4830, -1)			(5.8395, -1.9194, -1)			(4.7010, -1.6953, -1)		
	GCoD	0.6667			0.6976			0.7384		
	%	14%			19%			23%		
	ϵ_{90}	114.0191			102.4955			97.65193		
kC	β	(-1.0107, -1.0744, -1)			(-1.0107, -1.0744, -1)			(-0.8903, -1.0744, -1)		
	GCoD	0.5665			0.6135			0.6556		
	%	67%			67%			66%		
	ϵ_{90}	150.2014			150.2014			150.2834		
AkC	β	(-2.6754, -1.0658, -1)			(-2.7011, -0.9640, -1)			(-3.9149, -1.0070, -1)		
	GCoD	0.9901			0.9910			0.9915		
	%	69%			68%			69%		
	ϵ_{90}	150.0206			161.8515			155.8964		
MED	β	(-0.8019, -1.0319, -1)			(-2.6799, -1.0009, -1)			(-1.5141, -1.0345, -1)		
	GCoD	0.9911			0.9924			0.9928		
	%	74%			70%			70%		
	ϵ_{90}	155.184			157.4707			154.3846		

report: the maximum, minimum, median and mean width of the strips that are necessary to cover the 90% of the (validation) data for the seven runs.

From the above results, we note that the models that use vertical distance residuals need, in general larger strips to cover the 90% of the points. The strips are particularly large for the MEDIAN criterion, where the widest strips were obtained. This conclusion is justified since the quantile criteria accommodate a single point, but do not take into account the deviations to the remainder elements in the data (apart from the ordering in the residuals). Also, for the same reason, the conservative MAX criterion makes the models to require wider strips. The main observed difference between the MEDIAN and the MAX criteria is that whereas the behavior (in term of the fitting strips) of the MAX criterion is similar for the six choices of residuals, the MEDIAN gets very different results depending of the chosen residual. The most robust residuals, based on the smallest range between the maximum and minimum length of the strips, are the ℓ_1 , $\ell_{1.5}$, and ℓ_3 ; while with the same measure of robustness, the k -centrum criterion gets the best results.

To illustrate the quality of the optimal hyperplanes, in Figure 6 we show the values of the consumptions with respect to the actual consumptions for the first random fold in the experiments (in the validation sample that was not used to compute the hyperplanes).

TABLE 6. Results for Experiments for $d = 4$ and corrupting the X variables.

		V	ℓ_1	ℓ_∞
SUM	β	(8.7754, 0.2361, 0.1242, -0.0645, 1)	(-167.9861, 32.8678, -11.1472, -15.3593, 1)	(19.6624, 1.9411, 1.4336, -2.6949, 1)
	GCoD	0.0369	0.3527	0.7030
	%	8%	9%	15%
	ϵ_{90}	285.1339	172.616	166.2396
MAX	β	(11.2676, -0.8055, 0.4093, 0.3802, 1)	(95.4943, -2.3074, -2.7088, 4.5984, 1)	(76.9688, -2.1455, -2.9597, 4.6480, 1)
	GCoD	0.1200	0.5037	0.7852
	%	2%	9%	6%
	ϵ_{90}	243.9038	160.86	164.3572
SOS	β	(2.7637, 0.1306, 0.06391, -0.0111, 1)	(-35.0079, -17.4180, 5.1138, 8.8243, -1)	(14.4492, 2.3985, 1.8254, -3.4712, 1)
	GCoD	0.0409	0.5787	0.9085
	%	6%	9%	8%
	ϵ_{90}	285.0815	170.37	165.6255
1.5SUM	β	(3.1382, 0.1714, 0.0663, -0.03521)	(21.9152, -18.9245, 5.5144, 9.6284, -1)	(-20.1562, -2.0728, -1.5407, 2.9444, -1)
	GCoD	0.0418	0.4776	0.8349
	%	7%	8%	14%
	ϵ_{90}	282.7383	167.7096	165.9725
kC	β	(-6.8937, 0.1108, 0.0744, -0.0183, 1)	(-34.1432, -15.4977, 4.3066, 7.9523, -1)	(5.0421, 2.0898, 1.4381, -2.8638, 1)
	GCoD	0.0258	0.3487	0.6984
	%	8%	8%	15%
	ϵ_{90}	276.4327	168.3023	169.65
AkC	β	(-29.5486, 0.5489, 0.2119, 0.2342, 1)	(11.5813, 2.8055, -0.1579, 0.1805, 1)	(2.7269, 1.0225, 0.9985, 1.0072, 1)
	GCoD	0.1544	0.8716	0.9950
	%	12%	5%	82%
	ϵ_{90}	304.1316	306.9669	496.6216
MED	β	(11.3163, 0.5095, 0.5018, 0.0667, 1)	(15.2913, -1.38181, -0.1062, 9.6624, 1)	(2.3001, 1.0447, 1.0149, 1.0033, 1)
	GCoD	0.3706	0.8308	0.9941
	%	9%	11%	80%
	ϵ_{90}	283.331	251.5948	497.3323

		$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	β	(-25.3339, 7.2803, 0.3850, -6.5208, 1)	(-25.3339, 7.2803, 0.3850, -6.5208, 1)	(-48.9741, -2.5251, -1.5173, 3.4889, -1)
	GCoD	0.3973	0.4630	0.5446
	%	12%	12%	11%
	ϵ_{90}	167.1534	167.1534	163.8287
MAX	β	(-76.9688, 2.1455, 2.9597, -4.6480, -1)	(-76.9688, 2.1455, 2.9597, -4.6480, -1)	(-76.9688, 2.1455, 2.9597, -4.6480, -1)
	GCoD	0.5510345	0.6096547	0.677138
	%	6%	6%	6%
	ϵ_{90}	164.3572	164.3572	164.3572
SOS	β	(-19.8365, -24.1780, -1.6843, 23.0309, -1)	(-37.1798, -20.6518, -4.8914, 22.4924, -1)	(16.2930, 4.1351, 2.2042, -5.3890, 1)
	GCoD	0.6391	0.7149	0.7921
	%	9%	9%	4%
	ϵ_{90}	159.013	160.1321	165.3201
1.5SUM	β	(27.4692, 14.0582, 1.0081, -12.9659, 1)	(27.4555, 14.0608, 1.0082, -12.9683, 1)	(-20.4048, -3.2308, -1.6763, 4.1796, -1)
	GCoD	0.5314	0.6059	0.6909
	%	10%	10%	5%
	ϵ_{90}	162.8882	162.8875	164.1443
kC	β	(31.8219, 41.5015, -5.2288, -30.4070, 1)	(2.4227, 14.3655, 4.4768, -15.4827, 1)	(6.6713, -3.7849, -1.5627, 4.3751, -1)
	GCoD	0.3916	0.4629	0.5440
	%	5%	7%	4%
	ϵ_{90}	165.793	168.1855	165.9668
AkC	β	(7.9530, -1.6065, 0.3482, 0.8960, -1)	(-25.2618, -1.0371, -1.4553, 0.7368, -1)	(40.7617, -1.6662, -0.5106, 0.5624, -1)
	GCoD	0.7403	0.8148	0.8817
	%	7%	11%	9%
	ϵ_{90}	180.9401	244.0442	231.9954
MED	β	(-28.1536, -1.9062, -0.5785, 0.5246, -1)	(-51.5261, 1.9897, 1.0285, -0.5282, 1)	(6.9522, 1.2873, 1.0511, -0.1044, 1)
	GCoD	0.8278	0.8575	0.8941
	%	9%	8%	14%
	ϵ_{90}	237.8898	305.539	350.0691

The conclusions are that the vertical distance residuals do not fit well to the actual the trend of the validation data. The same conclusion also applies to the models that use the MAX criterion or ℓ_∞ residuals. On the other hand, the ℓ_1 -residual models seem to fit quite well to the data, whereas the ℓ_τ -residual models have similar (good) behavior. As expected the kC and AkC criteria, which are known to be very robust, actually capture the main information about the trend of the data.

6.2.2. Complete models. We also performed the same experiments for the whole data set. The three variables X_1 (incomes), X_2 (prices) and X_3 (consumptions) are now considered. The optimal hyperplanes are shown in Table 11 (since the coefficients are non zero they were divided by $-\beta_3$ to make easier the interpretation and representations of the models as $X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2$).

The summary of the results of the k -fold cross validation scheme (where the data set was partitioned exactly as in the bivariate case) is shown in Table 12. Finally, Figure 7 shows the values of the consumptions with respect to the actual consumptions for the first random fold in the experiments. From the results, one can observe that including all the variables in the model reduces the differences among the models obtained with the different methods. In this case, the consumption seems to be well linearly described by the incomes and prices. This conclusion is supported both by the projection and by the summary of k -cross validation experiments. The exceptionally bad performance of the MAX criterion in the former case (the model that only included X_1 and X_3), is now as good as the rest of the criteria. In addition, the inclusion of prices in the model fixes the, in most cases, senseless signs of the coefficients in the simple models in Table 9. One can observe that in those cases an increase of the incomes would predict a decrease of the consumptions. This unusual trend is fixed by introducing the prices in the complete model.

TABLE 7. Results for Experiments for $d = 4$ and corrupting the Y variables.

		V	ℓ_1	ℓ_∞
SUM	β	(1.9468, 0.9648, 0.9899, 1.0058, 1)	(-1.9158, -1.1083, -0.8751, -3.3186, -1)	(1.6655, -1.0083, -1.0530, -1.0446, -1)
	GCoD	0.5999	0.6538	0.9006
	%	78%	14%	76%
	ϵ_{90}	123.5456	149.6274	121.8106
MAX	β	(1 - 04.7766, -1.0780, -2.8506, -0.8355, -1)	(120.6153, -1.4207, -5.5268, -0.7782, -1)	(54.3395, 2.3207, 6.0411, 3.4977, 1)
	GCoD	0.3357	0.8267	0.9078
	%	12%	7%	12%
	ϵ_{90}	151.6067	147.4952	138.4277
SOS	β	(-12.1432, -0.8507, -1.0758, -1.1049, -1)	(25.1165, -1.2149, -5.4326, -1.1199, -1)	(-5.4787, -1.8048, -2.3397, -2.0389, -1)
	GCoD	0.4247	0.9015	0.9801
	%	45%	13%	15%
	ϵ_{90}	124.0456	135.9287	102.1587
1.5SUM	β	(-2.1265, -0.9557, -0.9984, -1.0235, -1)	(34.3751, -1.0783, -5.2458, -1.0619, -1)	(-0.6651, -1.3869, -1.5549, -1.5790, -1)
	GCoD	0.5106	0.8044	0.9485
	%	77%	11%	22%
	ϵ_{90}	124.3694	139.4734	95.54551
kC	β	(-0.3095, -0.9816, -1.0017, -1.009643, -1)	(2.1980, -0.8680, -0.9950, -3.4086, -1)	(-0.6929, -1.0211, -1.0606, -1.0666, -1)
	GCoD	0.5275	0.6525	0.8835
	%	80%	10%	74%
	ϵ_{90}	123.0891	145.6142	120.8033
AkC	β	(-7.2126, -0.9981, -1.2345, -0.9988, -1)	(-1.7307, -0.9801, -1.0396, -1.0121, -1)	(0.1128, -0.9847, -1.0149, -1.0013, -1)
	GCoD	0.8785	0.9933	0.9981
	%	57%	77%	80%
	ϵ_{90}	105.7586	120.4785	121.9634
MED	β	(-8.4437, -1.0328, -1.1891, -0.9958, -1)	(-3.0605, -0.9660, -1.0175, -1.0366, -1)	(-1.7471, -0.9713, -0.9881, -1.0144, -1)
	GCoD	0.9011	0.9921	0.9980
	%	58%	76%	79%
	ϵ_{90}	105.9371	123.0289	123.8959

		$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	β	(0.5934, -1.0202, -1.0588, -1.0264, -1)	(0.6616, -1.0203, -1.0584, -1.0270, -1)	(0.9775, -1.0098, -1.0563, -1.0343, -1)
	GCoD	0.7489	0.8006	0.8418
	%	80%	80%	78%
	ϵ_{90}	119.4431	119.5293	120.6788
MAX	β	(120.6153, -1.4207, -5.5268, -0.7782, -1)	(-54.3395, -2.3207, -6.0411, -3.4977, -1)	(-54.3395, -2.3207, -6.0411, -3.4977, -1)
	GCoD	0.8267	0.8384	0.8643
	%	7%	12%	12%
	ϵ_{90}	147.4952	138.4277	138.4277
SOS	β	(-14.4853, 1.5436, 4.4201, 1.5950, 1)	(-0.3904, 1.7361, 2.9264, 2.0617, 1)	(4.7620, 1.9721, 2.5444, 2.0415, 1)
	GCoD	0.9022	0.9272	0.9514
	%	13%	10%	12%
	ϵ_{90}	131.3351	114.7621	106.4697
1.5SUM	β	(15.7120, -1.1641, -2.6186, -1.8366, -1)	(-0.8627, -1.4497, -1.6239, -1.9098, -1)	(-0.6434, -1.4056, -1.5798, -1.5348, -1)
	GCoD	0.8079	0.8565	0.8965
	%	21%	22%	20%
	ϵ_{90}	114.939	97.67539	97.29497
kC	β	(-1.0976, -1.0234, -1.0643, -1.0656, -1)	(-1.0942, -1.0234, -1.0641, -1.0656, -1)	(-0.7613, -1.0216, -1.0617, -1.0665, -1)
	GCoD	0.7053	0.7661	0.8144
	%	74%	74%	74%
	ϵ_{90}	120.25	120.262	120.6901
AkC	β	(0.8072, -0.9319, -1.1111, -1.0901, -1)	(-1.5573, -0.9672, -0.9991, -1.0184, -1)	(2.4443, -1.0165, -0.9923, -1.0147, -1)
	GCoD	0.9929	0.9954	0.9930
	%	64%	77%	82%
	ϵ_{90}	124.0139	123.7847	123.5452
MED	β	(-0.6735, -0.9887, -1.0180, -0.9497, -1)	(0.4156, -0.9995, -1.0147, -1.0116, -1)	(-1.1572, -0.9753, -1.0309, -0.9853, -1)
	GCoD	0.9945	0.9949	0.9964
	%	75%	81%	78%
	ϵ_{90}	118.3319	121.9701	120.0091

TABLE 8. Estimations for the bidimensional Durbin-Watson's dataset.

	V	ℓ_1	ℓ_∞
SUM	(4.0898, -1.1454, -1)	(10.8840, -4.6184, -1)	(8.9764, -3.6797, -1)
MAX	(1.6986, -0.0196, -1)	(1.6986, -0.0196, -1)	(-0.5963, 1.1530, -1)
SOS	(2.9993, -0.6309, -1)	(13.5934, -6.0703, -1)	(7.0978, -2.7353, -1)
1.5SUM	(4.0730, -1.1566, -1)	(10.6113, -4.5067, -1)	(7.9926, -3.1851, -1)
kC	(5.5288, -1.9236, -1)	(8.7033, -3.5303, -1)	(7.6654, -2.9977, -1)
AkC	(2.7467, -0.4031, -1)	(17.1272, -7.6311, -1)	(18.4349, -8.2833, -1)
MED	(2.4167, -0.2310, -1)	(28.0156, -13.0469, -1)	(23.4462, -10.7748, -1)

	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	(10.8840, -4.6184, -1)	(10.8746, -4.6138, -1)	(9.8917, -4.1344, -1)
MAX	(1.6986, -0.0196, -1)	(-0.5963, 1.1530, -1)	(-0.5963, 1.1530, -1)
SOS	(13.1400, -5.8376, -1)	(10.9561, -4.7162, -1)	(8.7832, -3.6006, -1)
1.5SUM	(10.4466, -4.4233, -1)	(9.6868, -4.0399, -1)	(8.9821, -3.6851, -1)
kC	(8.0130, -3.1750, -1)	(8.0455, -3.1914, -1)	(8.5389, -3.4427, -1)
AkC	(13.9827, -6.0670, -1)	(21.0745, -9.6064, -1)	(20.6955, -9.4349, -1)
MED	(24.0656, -11.0819, -1)	(6.4510, -2.4601, -1)	(28.0150, -13.0466, -1)

TABLE 9. Marginal variations for each of the models.

	V	ℓ_1	ℓ_∞	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	-1.1455	0	-0.7863	-0.0464	-0.2070	-0.4395
MAX	-0.0196	-0.0196	0.5355	-0.0196	0.4949	0.5151
SOS	-0.6309	0	-0.7322	-0.0291	-0.2029	-0.4597
1.5SUM	-1.1566	0	-0.7610	-0.0505	-0.2332	-0.4564
kC	-1.9236	0	-0.7498	-0.0961	-0.2853	-0.4660
AkC	-0.4032	0	-0.8922	-0.0270	-0.1029	-0.3147
MED	-0.2310	0	-0.9150	-0.0081	-0.3488	-0.2711

FIGURE 5. Estimated lines for the data in [16] .

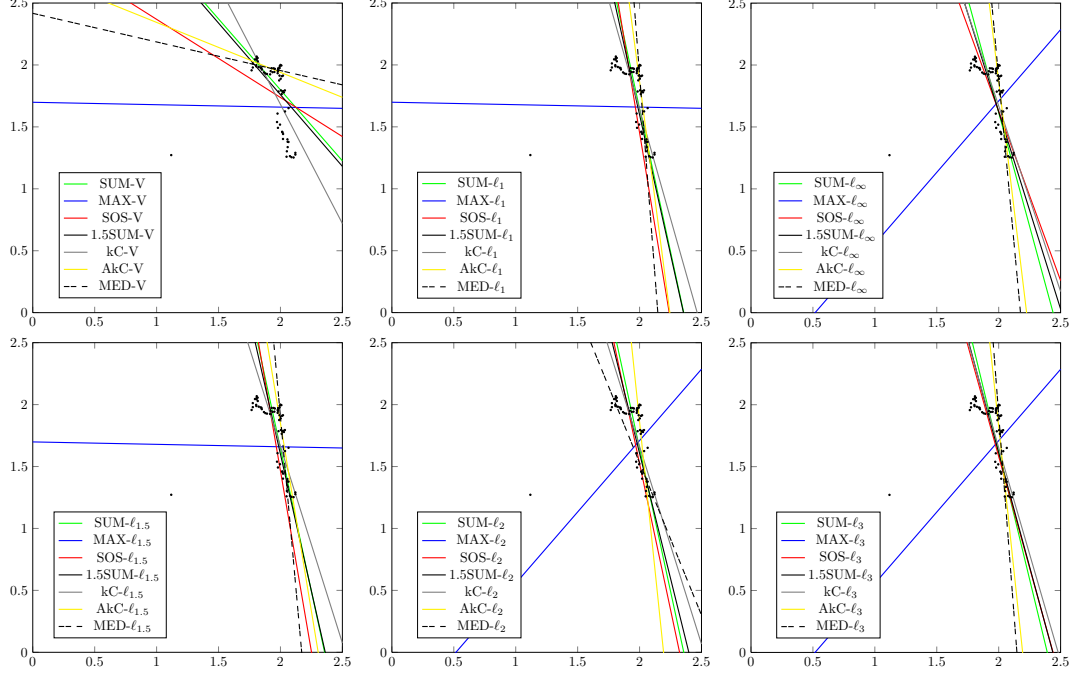


TABLE 10. Summary of k-fold cross validations experiments for the bidimensional Durbin-Watson's dataset.

		V	ℓ_1	ℓ_∞	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	min ε_{90}	0.1590	0.0560	0.0702	0.0491	0.0459	0.0560
	max ε_{90}	0.3049	0.1645	0.1444	0.1477	0.1480	0.1480
	median ε_{90}	0.2366	0.0983	0.0923	0.0881	0.0828	0.0983
	$\bar{\varepsilon}_{90}$	0.2330	0.1027	0.0982	0.0958	0.0959	0.1021
MAX	min ε_{90}	0.1262	0.1274	0.1262	0.1262	0.1262	0.1274
	max ε_{90}	0.3955	0.3955	0.3663	0.3663	0.3663	0.3955
	median ε_{90}	0.3664	0.3664	0.3621	0.3621	0.3621	0.3664
	$\bar{\varepsilon}_{90}$	0.3337	0.3338	0.3222	0.3222	0.3222	0.3338
SOS	min ε_{90}	0.1372	0.0844	0.0566	0.0568	0.0633	0.0793
	max ε_{90}	0.4072	0.1264	0.1163	0.1202	0.1235	0.1253
	median ε_{90}	0.2878	0.0962	0.0983	0.0879	0.0961	0.0961
	$\bar{\varepsilon}_{90}$	0.2980	0.1005	0.0973	0.0900	0.0905	0.0983
1.5SUM	min ε_{90}	0.1437	0.0476	0.0488	0.0524	0.0499	0.0478
	max ε_{90}	0.3091	0.1353	0.1199	0.1254	0.1308	0.1334
	median ε_{90}	0.2260	0.0834	0.0852	0.0910	0.0885	0.0841
	$\bar{\varepsilon}_{90}$	0.2349	0.0922	0.0872	0.0869	0.0884	0.0917
kC	min ε_{90}	0.1236	0.0414	0.0655	0.0495	0.0480	0.0412
	max ε_{90}	0.2843	0.1220	0.1147	0.1163	0.1185	0.1219
	median ε_{90}	0.1281	0.0837	0.0837	0.0851	0.0851	0.0855
	$\bar{\varepsilon}_{90}$	0.1511	0.0827	0.0834	0.0800	0.0809	0.0821
AkC	min ε_{90}	0.4482	0.0421	0.0429	0.0367	0.0892	0.0484
	max ε_{90}	0.6677	0.2039	0.1853	0.2122	0.4654	0.1981
	median ε_{90}	0.5162	0.1722	0.1296	0.1605	0.1534	0.1466
	$\bar{\varepsilon}_{90}$	0.5282	0.1434	0.1338	0.1417	0.1914	0.1373
MED	min ε_{90}	0.4275	0.1182	0.1147	0.0979	0.1182	0.0615
	max ε_{90}	0.6375	0.2170	0.4612	0.2203	0.2137	0.2101
	median ε_{90}	0.5503	0.1712	0.1761	0.1701	0.1393	0.1565
	$\bar{\varepsilon}_{90}$	0.5406	0.1651	0.2093	0.1614	0.1501	0.1478

TABLE 11. Estimations for the Durbin-Watson's dataset.

	V	ℓ_1	ℓ_∞
SUM	(4.4817, 0.0696, -1.3374, -1)	(4.555, 0.0587, -1.3623, -1)	(4.1367, 0.3502, -1.4305, -1)
MAX	(4.5227, 0.0646, -1.3519, -1)	(4.6159, -0.013, -1.3273, -1)	(4.1355, 0.5086, -1.5758, -1)
SOS	(3.9725, 0.0331, -1.0692, -1)	(4.404, 0.1369, -1.3881, -1)	(4.404, 0.1369, -1.3881, -1)
1.5SUM	(4.404, 0.1369, -1.3881, -1)	(4.404, 0.1369, -1.3881, -1)	(4.404, 0.1369, -1.3881, -1)
kC	(4.4159, 0.0288, -1.2753, -1)	(4.4905, 0.0635, -1.3425, -1)	(4.3334, 0.1325, -1.3317, -1)
AkC	(4.4355, 0.0655, -1.3183, -1)	(4.4521, 0.0585, -1.3197, -1)	(4.4688, 0.0535, -1.323, -1)
MED	(4.4288, 0.0488, -1.2979, -1)	(4.5075, 0.0634, -1.3476, -1)	(4.3559, 0.1431, -1.3489, -1)

	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	(4.4445, 0.0698, -1.3242, -1)	(4.472, 0.0633, -1.331, -1)	(4.4922, 0.0619, -1.3386, -1)
MAX	(4.4155, 0.0352, -1.2797, -1)	(4.3938, 0.1107, -1.3377, -1)	(4.2655, 0.1691, -1.3326, -1)
SOS	(4.3498, 0.1131, -1.3201, -1)	(4.3498, 0.1131, -1.3201, -1)	(4.3498, 0.1131, -1.3201, -1)
1.5SUM	(4.2123, 0.4308, -1.5386, -1)	(4.0853, 0.4429, -1.4891, -1)	(3.6048, 0.7761, -1.5744, -1)
kC	(5.2647, -0.6758, -1.0312, -1)	(3.5719, 1.1094, -1.8642, -1)	(3.4912, 1.0623, -1.7796, -1)
AkC	(4.1061, 0.5015, -1.551, -1)	(4.1579, 0.467, -1.5434, -1)	(4.2963, 0.3239, -1.4761, -1)
MED	(4.3576, 0.2689, -1.4559, -1)	(4.0772, 0.4066, -1.4415, -1)	(76.3635, 25.0913, -61.4268, -1)

FIGURE 6. Responses in the dependent variable by residuals for the bivariate case (SUM: red, MAX: blue, SOS: green, 1.5SUM: yellow, kC: black, AkC: orange, MEDIAN: gray) .

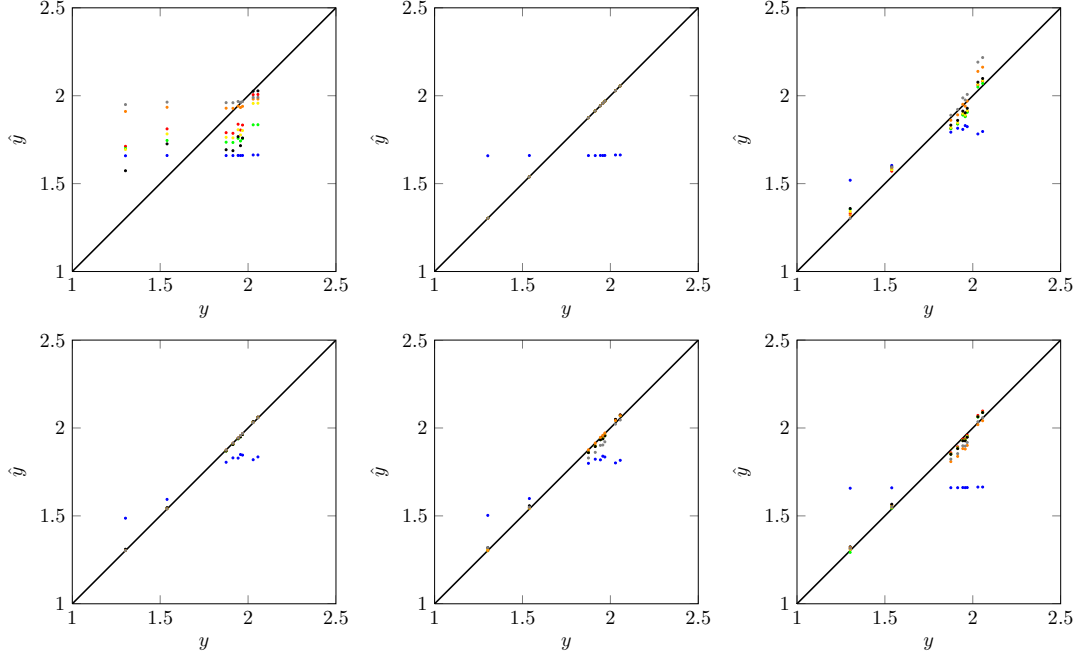


TABLE 12. Summary of k-fold cross validations experiments for the Durbin-Watson's dataset.

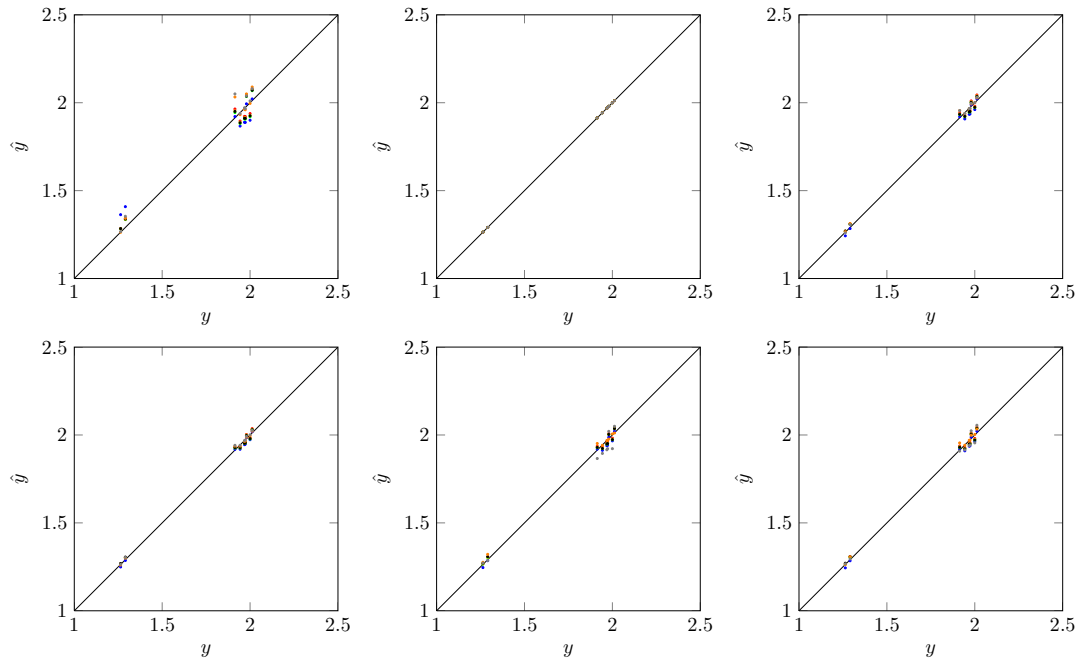
		V	ℓ_1	ℓ_∞	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	min ε_{90}	0.0369	0.0388	0.0315	0.0380	0.0346	0.0347
	max ε_{90}	0.0735	0.0741	0.0832	0.0743	0.0743	0.0732
	median ε_{90}	0.0629	0.0627	0.0647	0.0625	0.0625	0.0626
	ε_{90}	0.0573	0.0598	0.0616	0.0580	0.0567	0.0593
MAX	min ε_{90}	0.0562	0.0515	0.0515	0.0515	0.0515	0.0515
	max ε_{90}	0.0807	0.0762	0.0760	0.0760	0.0760	0.0762
	median ε_{90}	0.0701	0.0607	0.0644	0.0644	0.0607	0.0607
	ε_{90}	0.0678	0.0624	0.0641	0.0641	0.0624	0.0624
SOS	min ε_{90}	0.0255	0.0362	0.0310	0.0321	0.0327	0.0327
	max ε_{90}	0.0656	0.0683	0.0691	0.0678	0.0675	0.0675
	median ε_{90}	0.0586	0.0583	0.0568	0.0586	0.0581	0.0582
	ε_{90}	0.0547	0.0541	0.0537	0.0543	0.0528	0.0529
1.5SUM	min ε_{90}	0.0262	0.0342	0.0292	0.0308	0.0314	0.0316
	max ε_{90}	0.0685	0.0709	0.0713	0.0691	0.0703	0.0703
	median ε_{90}	0.0617	0.0563	0.0587	0.0559	0.0556	0.0558
	ε_{90}	0.0553	0.0547	0.0546	0.0527	0.0531	0.0532
kC	min ε_{90}	0.0269	0.0368	0.0265	0.0251	0.0272	0.0272
	max ε_{90}	0.0650	0.0700	0.0698	0.0709	0.0709	0.0700
	median ε_{90}	0.0588	0.0564	0.0559	0.0559	0.0569	0.0571
	ε_{90}	0.0514	0.0549	0.0536	0.0534	0.0538	0.0535
akC	min ε_{90}	0.0349	0.0338	0.0360	0.0305	0.0256	0.0604
	max ε_{90}	0.1042	0.1041	0.1017	0.3524	0.1100	0.1303
	median ε_{90}	0.0906	0.0888	0.0820	0.0885	0.0676	0.0931
	ε_{90}	0.0815	0.0799	0.0778	0.1115	0.0713	0.0923
MED	min ε_{90}	0.0342	0.0329	0.0346	0.0332	0.0429	0.0270
	max ε_{90}	0.1064	0.0994	0.0997	0.1102	0.3410	0.3266
	median ε_{90}	0.0709	0.0872	0.0894	0.0649	0.0844	0.0714
	ε_{90}	0.0738	0.0784	0.0794	0.0671	0.1215	0.1012

7. CONCLUSIONS AND FURTHER RESEARCH

This paper introduces a new framework for fitting hyperplanes to a given set of points by considering distance-based residuals and applying generalized ordered weighted averaging aggregation criteria. Mathematical programming formulations are proposed for those models and some properties are proven. Two important particular cases of residuals are analyzed in more detail, namely those induced by block norms or ℓ_τ norms for $\tau \geq 1$. A new goodness of fitting measure is also introduced for this framework, which extends the classical coefficient of determination in least sum of squares fitting with vertical distances. Extensive computational experiments run in Gurobi under R are reported in order to illustrate and validate the new methodology for computing optimal fitting hyperplanes.

The results in this paper admit some extensions applying similar tools. Among them we mention regularization adding constraints to overcome ill-posed data set, the simultaneous computation of several (more than one) hyperplanes to a given data set such that each single point is “allocated” to its *closest model*. This approach would allow to analyze structural changes on the behavior of the data (in different periods of time or for different values of one of the variables). The main, non trivial, difference between those models and the ones proposed in this paper is analogous to that that exists between the so-called single-facility and multifacility

FIGURE 7. Responses in the dependent variable by residuals for the $d = 3$ case (SUM: red, MAX: blue, SOS: green, 1.5SUM: yellow, kC: black, AkC: orange, MEDIAN: gray) .



location problems (see [32]). It is well-known that multifacility problems become easily hard even if the single-facility case were *easy*. Hence, although very interesting, the above extension needs further analysis. Another interesting extension is the use of mathematical programming tools to fit hyperplanes to binary data. The usual techniques to estimate those models are based on likelihood estimation since least squares estimation is known to get no desirable results on this type of data. Here our proposal will fit in a natural way and will deserve further attention.

ACKNOWLEDGEMENTS

The first and second authors were partially supported by the project MTM2016-74983-C2-1-R and MTM2013-46962-C2-1-P (MINECO, Spain).

REFERENCES

REFERENCES

- [1] Amaldi, E. and Coniglio, S., and Taccari, L. (2016). *Discrete optimization methods to fit piecewise affine models to data points*, Computers & Operations Research 75, 214–230.
- [2] Arthanary, T. S. and Dodge, Y. (1980). *Mathematical Programming in Statistics*, John Wiley and Sons.
- [3] Atkinson, A. C. and Cheng, T. C. (1999). *Computing least trimmed squares regression with the forward search*. Stat. Comp. 9, 251–263.
- [4] Balas, E. (1979). *Disjunctive Programming*. Ann. Discrete Math. 5, 3–51.
- [5] Bargiela, A, Hartley, J.K. (1993). *Orthogonal linear regression algorithm based on augmented matrix formulation*, Computers & Operations Research 20(8), 829–836.
- [6] Bertsimas, D. & Shioda, R. (2007). *Classification and Regression via Integer Optimization*. Oper. Res. 55(2): 252–271.
- [7] Bertsimas, D. & Mazumder, R. (2014) *Least Quantile regression via modern optimization*. Ann. Stat. 42 (6), 2494–2525.
- [8] Bertsimas, D., King, A. & Mazumder, R. (2016). *Best subset selection via a modern optimization lens*. Annals of Statistics 44 (2), 813–852.
- [9] Blanco V., Puerto J. and El-Haj Ben-Ali S. (2014). *Revisiting several problems and algorithms in continuous location with ℓ_r norms*. Comput. Optim. Appl. 58(3), 563–595.
- [10] Boggs, P. T., and J. E. Rogers (1990). *Orthogonal Distance Regression*, Contemp. Math. 112, 183–194.
- [11] Carrizosa, E., Conde, E., Fernández, F.R., Muñoz, M. and Puerto, J. (1995) *Pareto optimality in Linear Regression*. J. Math. Anal. Appl. 190, 129–141.
- [12] Carrizosa, E. and Plastria, F. (1995). *The determination of a “least quantile of squares regression line” for all quantiles*. Computational Statistics & Data Analysis, 20(5):467–479.
- [13] Cavalier, T., Melloy, B. (1991). *An Iterative Linear Programming Solution to the Eudidean Regression Model*, Comput. Oper. Res. 18 (8), 655–661.
- [14] Diaz-Báñez, J.M., Mesa, J.A, and Schöbel, A. (2004). *Continuous Location of dimensional structures*. European Journal of Operational Research 152 (1), 22–44.
- [15] Drezner, Z., Steiner, S. and Wesolowsky, G.O. (2002). *On the circle closest to a set of points*, Computers & Operations Research, 29(6) 637–650.
- [16] Durbin, J. and Watson, G.S. (1951). *Testing for serial correlation in least squares regression II*. Biometrika, 38, 159–178.

- [17] Fernández, E., Pozo, M.A., and Puerto, J. (2014). *Ordered weighted average combinatorial optimization: Formulations and their properties*. Discrete Appl. Math. 169, 97–118.
- [18] Fernández, E., Pozo, M.A., Puerto, J. and Scozzari, A. (2016) *Ordered Weighted Average Optimization in Multiobjective Spanning Tree Problems*. European Journal of Operational Research, to appear 2016.
- [19] Giloni, A. and Padberg, M. (2002). *Alternative methods of linear regression*, Math. Comput. Model., 35 (3–4), 361–374.
- [20] Grzybowski J, Nickel S, Pallaschke D, Urbański R (2011). *Ordered median functions and symmetries*. Optimization 60:801–811
- [21] Hoerl, A. and Kennard, R. (1988). *Ridge regression*. In Encyclopedia of Statistical Sciences, vol. 8, pp. 129–136. New York: Wiley.
- [22] Lee, S. and Grossmann, I. (2000). *New Algorithms for Nonlinear Generalized Disjunctive Programming*. Comput. Chem. Eng. 24, 2125–214.
- [23] Love, R.F. and Morris, J.G. (1972). *Modelling Inter-City Road Distances by Mathematical Functions*. Oper. Res. Q. 23 (1), 61–71.
- [24] Mangasarian, O.L. (1999). *Arbitrary-norm separating plane*. Oper. Res. Lett., 24 (1– 2):15–23.
- [25] Marín, A., Nickel, S., Puerto, J. and Velten, S. (2009) *A flexible model and efficient solution strategies for discrete location problems*. Discrete Applied Mathematics, 157(5): 1128–1145.
- [26] Megiddo, N. and Tamir, A (1983). *Finding least-distance lines*. SIAM J. on Algebraic and Discrete Methods, 4(2):207–211.
- [27] Miller, A. (2002). *Subset selection in regression*. CRC Press Washington.
- [28] McKean, JW and Sievers GL (1987). *Coefficients of determination for least absolute deviation analysis*, Stat. Probabil. Lett. 5(1), 49–54
- [29] Miyashiro, R and Takano, Y (2015). *Mixed integer second-order cone programming formulations for variable selection in linear regression*. European Journal of Operational Research 247(3), 721–731.
- [30] Narula, SC and Wellington JF (2007). *Multiple criteria linear regression*, European Journal of Operational Research 181(2) , 767–772.
- [31] Nickel S. and J. Puerto (1999). *A unified approach to network location*. Networks vol. 34, 283–290.
- [32] Nickel, S. and Puerto, J. (2005). *Facility Location - A Unified Approach*. Springer Verlag.
- [33] Pham-Gia, T. and Hung, T.L. (2001). *The mean and median absolute deviations*. Math. Comput. Model., 34, 921–936.
- [34] Pinson, P, Nielsen, H, Madsen, H and Nielsen, T. (2008). *Local linear regression with adaptive orthogonal fitting for the wind power application*, Stat. Comput. 58 (1), 59–71.
- [35] Rousseeuw, P. J. (1983). *Multivariate Estimation With High Breakdown Point*. Math. Stat. App. B, (Ed. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz), 283–297.
- [36] Rousseeuw, P. (1984), *Least median of squares regression*. J. Am. Stat. Assoc., 79, 871–880
- [37] Rousseeuw, P. and Leroy, A. *Robust Regression and Outlier Detection*. New York: Wiley, 2003.
- [38] Schöbel, A (1996). *Locating least-distant lines with block norms*. Studies in Locational Analysis 10, 139–150.
- [39] Schöbel, A (1997). *Locating line segments with vertical distances*. Studies in Locational Analysis 11, 143–158.
- [40] Schöbel, A (1998). *Locating least distant lines in the plane*. European Journal of Operational Research 106(1), 152–159.
- [41] Schöbel, A. (1999). *Locating Lines and Hyperplanes: Theory and Algorithms*. Kluwer Academic Publishers, vol. 25. ISBN: 9781461374282.
- [42] Humphreys, R. M. (1978). *Studies of Luminous Stars in Nearby Galaxies. I. Supergiants and O Stars in the Milky Way*, Astrophys. J. Suppl. S. , 38, 309–350
- [43] Stone, M. (1974). *Cross-Validatory Choice and Assessment of Statistical Predictions*, J. R. Stat. Soc. B 36, 111–147.
- [44] Thoi, R. (1999). *D.C. programming: An overview*. J. Optimiz. Theory App. , 193(1), 1–43.
- [45] Van Huffel, S. and Vanderwalle, J. (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM Frontiers in Applied Mathematics.
- [46] Ward, J. E. and Wendell, R. E. (1980). *A new norm for measuring distance which yields linear location models*, Oper. Res. 28, 836–844.
- [47] Ward, J. E. and Wendell, R. E. (1985). *Using block norms for location modeling*, Oper. Res., 33, 1074–1090
- [48] Yager, R.R. and Beliakov, G. (2010), *OWA Operators in Regression Problems*, IEEE T. Fuzzy Syst. 18 (1), 106–113.

DEPT. QUANTITATIVE METHODS FOR ECONOMICS & BUSINESS, UNIVERSIDAD DE GRANADA.
E-mail address: `vblanco@ugr.es`

DEPT. ESTADÍSTICA E INVESTIGACIÓN OPERATIVA, UNIVERSIDAD DE SEVILLA.
E-mail address: `puerto@us.es`

DEPT. QUANTITATIVE METHODS FOR ECONOMICS & BUSINESS, UNIVERSIDAD DE GRANADA.
E-mail address: `romansg@ugr.es`